On extended boundary sequences of morphic and Sturmian words

Michel Rigo   Manon Stipulanti   Markus A. Whiteland

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 $n + \ell$ ($n \geqslant \ell \geqslant 1$). 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 abstract numeration systems S, we show that if an infinite word is S-automatic, then the same holds for its ℓ-boundary sequence. In particular, they are both morphic (or generated by an HD0L system). To precise the limits of this result, we discuss examples of non-addable numeration systems and S-automatic words for which the boundary sequence is nevertheless S-automatic and conversely, S-automatic words with a boundary sequence that is not S-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.

Mathematics Subject Classifications: 68R15, 11B85, 68V15

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 $n \geqslant \ell$. 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.

Department of Mathematics, University of Liège, Liège, Belgium (m.rigo@uliege.be, m.stipulanti@uliege.be, mwhiteland@uliege.be).
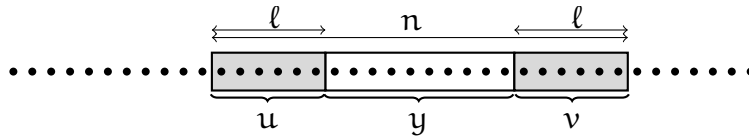
Figure 1: A sliding window where we focus on two regions of a fixed length.

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.

Definition 1. Let $\ell \in \mathbb{N}_{>0}$ and $\mathbf{x} \in A^{\mathbb{N}}$. For $n \geqslant \ell$, 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 $\partial_{\mathbf{x},\ell} := (\partial_{\mathbf{x},\ell}[n])_{n \geqslant \ell}$ 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 2. Consider the Fibonacci word $\mathbf{f} = 0100101001 \cdots$; the fixed point of the morphism $0 \mapsto 01, 1 \mapsto 0$. We have

$$\partial_{\mathbf{f}} = \mathsf{a\,b\,b\,a\,b\,b\,b\,a\,b\,b\,a\,b\,b\,b\,b\,a\,b\,b\,b\,b\,a\,b\,b\,a\,b\,b\,b\,b} \cdots,$$

where $\mathsf{a} := \{(0, 0), (0, 1), (1, 0)\}$ and $\mathsf{b} := \{0, 1\} \times \{0, 1\}$. For instance, $\partial_{\mathbf{f}}[1] = \mathsf{a}$ because the length-2 factors of $\mathbf{f}$ are $00, 01, 10$, while $\partial_{\mathbf{f}}[2] = \mathsf{b}$ 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

$$\partial_{\mathbf{f},2} = \mathsf{a\ b\ c\ d\ e\ f\ b\ c\ d\ b\ c\ d\ e\ f\ b\ c\ d\ e\ f\ b\ c\ d\ b\ c\ d} \cdots$$

where

$$
\begin{aligned}
\mathsf{a} &:= \{(00, 10), (01, 00), (01, 01), (10, 01), (10, 10)\}, \\
\mathsf{b} &:= \{(00, 00), (00, 01), (01, 01), (01, 10), (10, 00), (10, 10)\}, \\
\mathsf{c} &:= \{(00, 01), (00, 10), (01, 00), (01, 10), (10, 00), (10, 01)\}, \\
\mathsf{d} &:= \{(00, 00), (00, 10), (01, 00), (01, 01), (10, 01), (10, 10)\}, \\
\mathsf{e} &:= \{(00, 01), (01, 01), (01, 10), (10, 00), (10, 01), (10, 10)\}, \\
\mathsf{f} &:= \{(00, 10), (01, 00), (01, 01), (01, 10), (10, 01), (10, 10)\}.
\end{aligned}
$$

The first element $\partial_{\mathbf{f},2}[2] = \mathsf{a}$ is peculiar; it corresponds exactly to the five length-4 factors occurring in $\mathbf{f}$. Our Proposition 45 shows that a appears only once in $\partial_{\mathbf{f}}$. Then, e.g., $\partial_{\mathbf{f},2}[3] = \mathsf{b}$ 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 $u \in \{0, 1\} \times \{0, 1\}$. All letters, except a, appear infinitely often in $\partial_{\mathbf{f},2}$: see Theorem 37.

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 [19]. 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 [42], 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 [32]. 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, [23]. 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 [12] and was further studied in [24], where it is shown that the boundary sequence of a k-*automatic* word (in the sense of Allouche and Shallit [2]: see Definition 9) 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 [13] (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 [16, 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 [46]). For instance, computing the 2-binomial complexity of generalized Thue–Morse words [31] 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 [49]; there some general properties of boundary sequences of binary words are required (see [49, Lem. 4.6]). 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 [12]. Guo, Lü, and Wen combine this result with theirs in [24] to establish a large family of infinite words with automatic abelian complexity.

Let $k \geqslant 1$. 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 [27]. 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 $u \equiv_1 v$ [27,

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 [40, Thm. 3.6] (compare to Proposition 45). 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 [40, 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 [47]. With Theorem 16, we prove that for a large class of numeration systems S, if $\mathbf{x}$ is an S-automatic word, then the boundary sequence $\partial_{\mathbf{x}}$ is again S-automatic. Our approach generalizes the arguments provided by [24]. 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 for which the graph of addition is regular (see Definition 6 for details).

As an alternative, we observe that a classical effective procedure (Theorem 13) transforming formulae to automata can be extended to addable abstract numeration systems S. The S-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 comparisons and indexing into an S-automatic sequence.

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 finiteness of a suitable kernel captures all morphic words (see Theorem 11). To identify the contours of our result, we also discuss the case where $\mathbf{x}$ is S-automatic and $\partial_{\mathbf{x}}$ is not S-automatic. To construct such examples, we have to consider non-addable numeration systems in Section 3.3.

We then turn to the other class of words under study. Letting $\mathbf{s}$ be a Sturmian word with slope α, with Theorem 37 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 47.

This paper is a long version of [48] presented at MFCS 2022. It contains many proofs (omitted due to space limitation) and, in particular, discussions about Sturmian words. This extended version includes work through examples using `Walnut`. In Section 2.3 we

explicitly compute the 2-boundary sequence of the Thue–Morse and Fibonacci words, see Examples 14 and 15. In Section 3.2, we present several examples of automatic sequences built on intrinsically non-addable numeration systems for which the boundary sequence is still automatic, see Propositions 19 and 23. Finally, the proof of Theorem 16 has been strengthened to a larger setting to include addable abstract numeration systems. This slightly broadens the presentation of the paper which is not limited to positional numeration systems anymore.

2 Preliminaries

Throughout this paper we let A denote a finite alphabet. Then A^n denotes the set of length-n words and $A^{\mathbb{N}}$ denotes the set of infinite words. Infinite words will usually, but not always, be indexed starting from 0. They will also be written in bold. For a finite word u, we let u^{ω} denote the concatenation of infinitely many copies of the word u, i.e., $u^{\omega} = uuu \cdots$. For two words u, v for which $w = uv$, we let wv^{-1} denote the prefix u and $u^{-1}w$ the suffix v. For a finite or infinite word $\mathbf{x}$, we let $\mathbf{x}[n]$ denote the letter at index n (assuming it is well-defined for this value of n, e.g., if $\mathbf{x}$ is a ℓ-boundary sequence, $n \geqslant \ell$). Similarly, for $m \geqslant n$ we set $\mathbf{x}[n, m] := \mathbf{x}[n] \cdots [m]$. For any integer $n \geqslant 0$, we let $\mathrm{Fac}_n(\mathbf{x})$ denote the set of length-n factors of $\mathbf{x}$; we write $\mathrm{Fac}(\mathbf{x}) = \cup_{n \geqslant 0} \mathrm{Fac}_n(\mathbf{x})$. A factor u of an infinite word $\mathbf{x} \in A^{\mathbb{N}}$ is called *right special* if there exist distinct letters $a, b \in A$ such that $ua, ub \in \mathrm{Fac}(\mathbf{x})$. We note that an infinite word $\mathbf{x}$ is aperiodic if and only if it has a right special factor for each length. For general references on numeration systems, see [21] and [7, Chap. 1–3]. We assume that the reader has some knowledge in automata theory. For a reference see [51] or [45, Chap. 1].

2.1 Basic properties of boundary sequences

Recall that in our definition of the boundary sequence, we inspect factors of length $n + \ell$ with $n \geqslant \ell$. This implies that the prefix and suffix of length ℓ forming the boundary pair do not overlap. The following observation justifies this choice in a sense.

Proposition 3. *Let $\mathbf{x}$ be an aperiodic word and $\ell \geqslant 1$ be an integer. Then the boundary set $\partial_{\mathbf{x}, \ell}[m]$, with $0 \leqslant m < \ell$, appears exactly once in the sequence $(\partial_{\mathbf{x}, \ell}[n])_{n \geqslant 0}$.*

Proof. Fix an integer m with $0 \leqslant m < \ell$. We show that $\partial_{\mathbf{x}, \ell}[m] \neq \partial_{\mathbf{x}, \ell}[n]$ for any $n > m$. The claim follows straightforwardly from this observation. We first observe that any boundary pair $(u_1 \cdots u_\ell, v_1 \cdots v_\ell)$ in $\partial_{\mathbf{x}, \ell}[m]$ satisfies $u_{m+1} \cdots u_\ell = v_1 \cdots v_{\ell-m}$. In particular, $u_\ell = v_{\ell-m}$ for any pair in $\partial_{\mathbf{x}, \ell}[m]$. Consider then the boundary set $\partial_{\mathbf{x}, \ell}[n]$ with $n > m$. Let $x = x_1 \cdots x_{n+\ell}$ be a factor of length $n + \ell$ such that $x_1 \cdots x_{n+\ell-m-1}$ is right special and $x_\ell \neq x_{n+\ell-m}$ (here $\ell < n + \ell - m$ so such a choice can be made). Now x defines the boundary pair $(x_1 \cdots x_\ell, x_{n+1} \cdots x_{n+\ell}) = (u_1 \cdots u_\ell, v_1 \cdots v_\ell)$ for which $u_\ell \neq v_{\ell-m}$, which shows that this pair cannot appear in $\partial_{\mathbf{x}, \ell}[m]$. This concludes the proof. $\square$

The above proposition is tight in the sense that there exist aperiodic words for which the boundary set $\partial_{\mathbf{x},\ell}[\ell]$ appears infinitely often in the boundary sequence $\partial_{\mathbf{x},\ell}$. This can be seen, e.g., from Proposition 45. Another quick example for this is the Champernowne word $\mathbf{c} = 0\,1\,00\,01\,10\,11\cdots$ (the concatenation of the radix-ordered binary representations of the naturals) for which $\partial_{\mathbf{c},\ell} = (\{0,1\}^\ell \times \{0,1\}^\ell)^\omega$.

Lemma 4. *For any $\ell \geqslant 1$, the ℓ-boundary sequence of an eventually periodic word is eventually periodic.*

Proof. Let $\mathbf{x} = uv^\omega$. We claim that $\partial_{\mathbf{x},\ell}[n + |v|] = \partial_{\mathbf{x},\ell}[n]$ for all $n \geqslant \max\{\ell, |u|\}$. Indeed, consider a factor x of length $n + |v| + \ell$ occurring at position i. We may write $x = x's$ with $|x'| = n + |v|$ and $|s| = \ell$. Since $n \geqslant |u|$, there exists a factorization $v = v_1 v_2$ such that x' ends with vv_1, and s is a prefix of $(v_2 v_1)^\omega$. The factor of length $n + \ell$ occurring at position i is thus $x'(v_2 v_1)^{-1}s$. We have shown that the boundary pairs $(\mathbf{x}[i, i+\ell-1], \mathbf{x}[i+n, i+n+\ell-1])$ and $(\mathbf{x}[i, i+\ell-1], \mathbf{x}[i+n+|v|, i+n+|v|+\ell-1])$ are equal. This suffices for the proof. $\qquad\square$

2.2 Numeration systems and automatic words

For general references about automatic words and abstract numeration systems, see [2] and [47] or [7, Chap. 3]. An *abstract numeration system* (ANS) is a triple $S = (L, A, <)$ with L an infinite regular language over the totally ordered alphabet A (with $<$). We say that L is the *numeration language*. 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, called the *(e)valuation map*, is denoted by val_S.

Example 5. Consider the ANS S built on the language $\alpha^*\beta^*$ over the ordered alphabet $\{\alpha < \beta\}$. The first few words in the language are $\varepsilon, \alpha, \beta, \alpha\alpha, \ldots$. Hence, $\mathrm{rep}_S(3) = \alpha\alpha$ and $\mathrm{val}_S(\alpha\alpha) = 3$.

In the following, we refer to the terminology introduced in [39] (addable systems are called regular in [52]). It is convenient to introduce a new padding symbol $\#$ which does not belong to the alphabet A. We let $A_\#$ denote the set $A \cup \{\#\}$. We extend the evaluation map to $\#^*L$ by setting $\mathrm{val}_S(\#^n w) = \mathrm{val}_S(w)$ for all $w \in L$ and $n \in \mathbb{N}$.

Definition 6. An abstract numeration system $S = (L, A, <)$ is *addable* if the following *graph of addition*, denoted by $\mathcal{L}_+$, is regular:

$$\left\{ \begin{pmatrix} u \\ v \\ w \end{pmatrix} \in (\#^*L)^3 \cap (A_\# \times A_\# \times A_\#)^* \mid \mathrm{val}_S(u) + \mathrm{val}_S(v) = \mathrm{val}_S(w) \right\} \setminus \begin{pmatrix} \# \\ \# \\ \# \end{pmatrix} (A_\# \times A_\# \times A_\#)^*.$$

Notice that words in the numeration language L do not start with $\#$; however, when dealing with tuples of such words, shorter S-representations are padded with leading $\#$'s to get words of equal length (so they can be processed by an automaton reading tuples of letters). Continuing Example 5, for instance, the triplet $\begin{pmatrix} \#\alpha \\ \#\beta \\ \alpha\alpha \end{pmatrix}$ belongs to $\mathcal{L}_+$.

Remark 7. Positional numeration systems (whose numeration language is regular) are special instances of ANS. Let us recall this classical setting. Let $U = (U_n)_{n \geqslant 0}$ 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 [21, 45], then for all $j \leqslant t$ 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 $\sup_{i \geqslant 0}(U_{i+1}/U_i)$ 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*. Any such system for which the numeration language $\operatorname{rep}_U(\mathbb{N})$ is regular is an ANS. For a positional numeration system, the existence of the digit 0 permits to avoid the introduction of an extra symbol #. Padding can thus be achieved using leading zeroes.

Example 8. In this example, the numeration system has no digit 0 and has the property of being unambiguous. Consider the ANS S built on the language $L = \{1, 2\}^*$ and the sequence $U = (2^n)_{n \geqslant 0}$. The first few words in L are $\varepsilon, 1, 2, 11, 12, 21, 22, \ldots$. The nth word $d_k \cdots d_0$ in L verifies $n = \sum_{i=0}^{k} d_i 2^i$, but the greedy U-representation of n is just its base-2 expansion over $\{0, 1\}$ and is therefore not equal to $\operatorname{rep}_S(n) \in \{1, 2\}^*$. The ANS S is not, strictly speaking, a positional numeration system. Nevertheless the graph of addition for triplets of S-representations is regular. In Fig. 2, which depicts a DFA accepting the corresponding language reading least significant digit first, the digits are processed from right to left. One simply has to deal with a carry $0, 1, 2$ stored within the state. Transitions are of the form $m \longrightarrow n$ with label $\left(\begin{smallmatrix} p \\ q \\ r \end{smallmatrix}\right)$, for states $m, n \in \{0, 1, 2\}$ and letters $p, q, r \in \{\#, 1, 2\}$, if and only if

$$m + p + q = r + 2n$$

where # is interpreted as 0. This is therefore an example of an addable ANS which is not a positional numeration system handling greedy expansions.

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 9. Let $S = (L, A, <)$ be an ANS. 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 an integer $k \geqslant 2$, if $\mathcal{A}$ is fed with the genealogically ordered language $L = \{\varepsilon\} \cup \{1, \ldots, k-1\}\{0, \ldots, k-1\}^*$, then $\mathbf{x}$ is said to be k-*automatic*. If $\mathcal{A}$ is fed with the U-representations of integers, with U a positional numeration system, $\mathbf{x}$ is said to be U-automatic.

Theorem 10 ([47]). *A word $\mathbf{x}$ is morphic if and only if it is S-automatic for some abstract numeration system S.*

$$\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}2\\1\end{pmatrix}, \begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}\#\\2\\1\end{pmatrix}, \begin{pmatrix}2\\\#\\1\end{pmatrix}$$

$$\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}\#\\i\\i\end{pmatrix}, \begin{pmatrix}i\\\#\\i\end{pmatrix} \qquad\qquad \begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}2\\1\end{pmatrix}, \begin{pmatrix}2\\2\end{pmatrix}$$

$$\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}2\\1\end{pmatrix}, \begin{pmatrix}2\\2\end{pmatrix}$$

DFA diagram with states 0, 1, 2. Transitions labelled:
$0 \to 1$: $\begin{pmatrix}\#\\\#\\1\end{pmatrix}, \begin{pmatrix}1\\\#\\2\end{pmatrix}, \begin{pmatrix}\#\\1\\2\end{pmatrix}$; $1 \to 2$: $\begin{pmatrix}i\\\#\\i\end{pmatrix}, \begin{pmatrix}2\\2\\\#\end{pmatrix}, \begin{pmatrix}1\\1\\2\end{pmatrix}$; lower arc $\begin{pmatrix}\#\\\#\\2\end{pmatrix}$
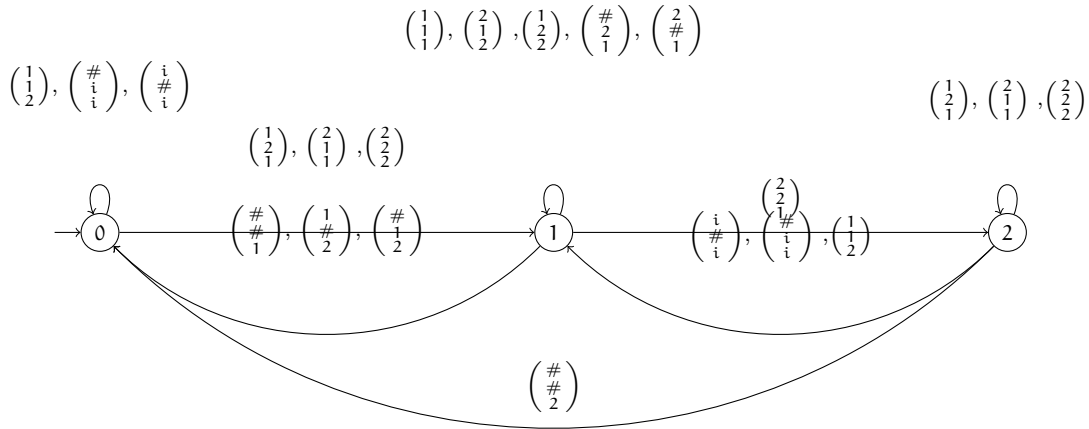
Figure 2: A DFA accepting $\mathcal{L}_+$ reading words from right to left (where $i \in \{1, 2\}$).

We note that the proof of the above theorem shows that the equivalence is completely effective: given the morphisms producing the word $\mathbf{x}$, one can construct an ANS S and a DFAO generating $\mathbf{x}$, and vice versa.

Fix $s \in A^*$. For a word $\mathbf{x}$, define the subsequence $\mathbf{x} \circ s$ by $(\mathbf{x} \circ s)[n] := \mathbf{x}[\mathrm{val}_S(p_{s,n} s)]$, 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 [7, Prop. 3.4.12–16].

Theorem 11 ([47]). *A word $\mathbf{x}$ is S-automatic if and only if its S-kernel is finite.*

Again, the theorem is completely effective: with the underlying ANS S fixed, given (a Turing machine generating) $\mathbf{x}$, and (the cardinality of) the S-kernel, one can compute the DFAO generating $\mathbf{x}$, and vice versa.

Example 12. Consider the Fibonacci numeration system based on the sequence of Fibonacci numbers $(F_n)_{n \geqslant 0}$ with $F_0 = 1$, $F_1 = 2$, and $F_{n+1} = F_n + F_{n-1}$ for $n \geqslant 1$. 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.

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.

In Section 3.1, we require S to be addable. Note that these assumptions of having a numeration language that is regular and addable are shared by many classical systems.

s	$(\mu_s(n))_{n \geqslant 0}$	s	$(\mu_s(n))_{n \geqslant 0}$
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, ...

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

For instance, the usual integer base numeration systems or the Fibonacci numeration system have all the assumed properties. For the latter system, the minimal automaton (reading most significant digits first) of $\mathcal{L}_+$ has 17 states (its transition table is given in [35]). The one reading least significant digits first has 22 states. The largest known family of positional systems with all these properties (addable with $\text{rep}_U(\mathbb{N})$ being regular) is the one of those based on a linear recurrence sequence whose characteristic polynomial is the minimal polynomial of a Pisot number [8, 45]. One practical difficulty when one wants to use automatic provers (such as `Walnut` [34]) is to be able to provide the relevant automaton for addition.

2.3 Link with first-order logic

The result stated below is at the origin of `Walnut`. It relies on the effective transformation of formulae to automata. It was first stated for integer-based systems. Making use of [11, Lem. 37 and Thm. 55], it was extended to addable systems:

Theorem 13 ([52, Thm. 6.4.1])**.** *Let S be an addable numeration system. There is an algorithm that, given a formula φ with no free variables, phrased in first-order logic, using only the universal and existential quantifiers, addition and subtraction of variables and constants, logical operations, comparisons, and indexing into a given S-automatic sequence $\mathbf{x}$, will decide the truth of the formula φ. Furthermore, if φ has $t \geqslant 1$ free variables, the algorithm produces a DFA that recognizes the language of all representations of t-tuples of natural numbers that make φ evaluate to true.*

As already mentioned in [24], 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 S: let $\mathbf{x} \in B^{\mathbb{N}}$ be S-automatic for an addable system S. 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 13 to conclude that $\partial_{\mathbf{x},\ell}$ is S-automatic. These arguments appear in [52, §8.1.11] in the case $\ell = 1$.

Example 14. We use the strategy described in [52, Sec. 8.1.11], where the boundary sequence of the Fibonacci word is computed. Here, we consider the Thue–Morse word $\mathbf{t}$ and show how to get its 2-boundary sequence

$$\partial_{\mathbf{t},2} = \mathtt{abcdedfdgdgdgdcdgdgdgdgdgdgdgdfdgd}\cdots$$

using `Walnut`. Let $\alpha, \beta, \gamma, \delta \in \{0, 1\}$. If there exists some position i such that

$$(\mathbf{t}[i]\mathbf{t}[i+1], \mathbf{t}[i+n]\mathbf{t}[i+n+1]) = (\alpha\beta, \gamma\delta),$$

then this pair belongs to $\partial_{\mathbf{t},2}[n]$. In `Walnut`, depending on the value of $\alpha, \beta, \gamma, \delta$, we provide sixteen definitions of the form

```
def TMboundαβγδ "Ei T[i]=@α & T[i+1]=@β & T[i+n]=@γ & T[i+n+1]=@δ";
```

which create deterministic automata recognizing base-2 expansions of the sets

$$T_{\alpha\beta,\gamma\delta} = \{n \in \mathbb{N} : (\alpha\beta, \gamma\delta) \in \partial_{\mathbf{t},2}[n]\}.$$

In particular, `$TMboundαβγδ(n)` evaluates to `TRUE` whenever n belongs to $T_{\alpha\beta,\gamma\delta}$. A direct inspection shows that only seven different 2-boundary sets occur in $\partial_{\mathbf{t},2}$:

$$a := \{0,1\}^2 \times \{0,1\}^2 \setminus \{(00,00),(00,01),(01,11),(10,00),(11,10),(11,11)\},$$
$$b := \{0,1\}^2 \times \{0,1\}^2 \setminus \{(00,00),(00,11),(01,10),(10,01),(11,00),(11,11)\},$$
$$c := \{0,1\}^2 \times \{0,1\}^2 \setminus \{(00,10),(01,00),(10,11),(11,01)\},$$
$$d := \{0,1\}^2 \times \{0,1\}^2 \setminus \{(00,00),(00,11),(11,00),(11,11)\},$$
$$e := \{0,1\}^2 \times \{0,1\}^2 \setminus \{(00,11),(11,00)\},$$
$$f := \{0,1\}^2 \times \{0,1\}^2 \setminus \{(00,01),(01,11),(10,00),(11,10)\},$$
$$g := \{0,1\}^2 \times \{0,1\}^2.$$

This can be checked as follows. For the set a, we provide the following definition

```
def TMbounda "~$TMbound0000(n) & ~$TMbound0001(n) & $TMbound0010(n)
            & $TMbound0011(n) & $TMbound0100(n) & $TMbound0101(n)
            & $TMbound0110(n) & ~$TMbound0111(n) & ~$TMbound1000(n)
            & $TMbound1001(n) & $TMbound1010(n) & $TMbound1011(n)
            & $TMbound1100(n) & $TMbound1101(n) & ~$TMbound1110(n)
            & ~$TMbound1111(n)";
```

where `$TMbound0(n)` evaluates to `TRUE` whenever $\partial_{\mathbf{t},2}[n] = a$. Similar definitions are readily written for $b, \ldots, g$. The following expression evaluates to `TRUE`

```
eval TM2boundarycheck "An (n>1) => (($TMbounda(n) | $TMboundb(n)
                        | $TMboundc(n) | $TMboundd(n) | $TMbounde(n)
                        | $TMboundf(n) | $TMboundg(n)))":
```

meaning that we are not missing any 2-boundary set. We combine the seven automata produced by `Walnut` accepting base-2 expansions of the integers n such that $\partial_{t,2}[n]$ is a particular letter in $\{a, \ldots, g\}$ into the DFAO depicted in Fig. 3 using the command

```
combine TM2boundarySequence TMbounda TMboundb TMboundc
                            TMboundd TMbounde TMboundf TMboundg:
```
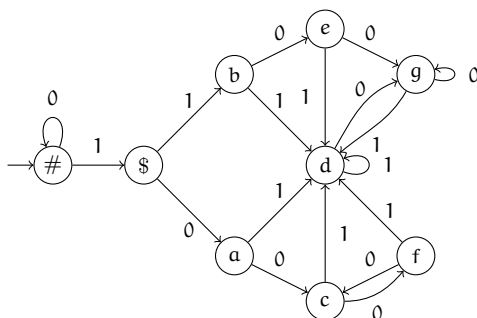

Figure 3: A DFAO producing $\partial_{t,2}$.

Inspecting Fig. 3, we notice that a, b and e appear exactly once. On the other hand, we have $\partial_{t,2}[n] = c$ precisely when n is an even power of 2, while $\partial_{t,2}[n] = f$ if and only if n is an odd power of 2 strictly larger than 2. Finally $\partial_{t,2}[n] = d$ precisely when n is odd and at least 5.

Through the effective conversion of the automaton to a morphic representation of the word, we get the 2-uniform morphism generating the 2-boundary sequence (prepended with two symbols #\$, because the 2-boundary sequence is indexed starting at 2):

$$\begin{cases} \# & \mapsto \#\$ \\ \$ & \mapsto ab \\ a & \mapsto cd \end{cases} \qquad \begin{cases} b & \mapsto ed \\ c & \mapsto fd \\ d & \mapsto gd \end{cases} \qquad \begin{cases} e & \mapsto gd \\ f & \mapsto cd \\ g & \mapsto gd. \end{cases}$$

Example 15. For the Fibonacci word $\mathbf{f}$, the strategy is similar to the one given in the previous example. Instead of binary expansions, we simply make use of Fibonacci expansions which are also available in `Walnut`. For instance

```
def Fbound00 "?msd_fib Ei F[i]=@0 & F[i+n]=@0";
```

is such that `Fbound00(n)` evaluates to `TRUE` whenever $(0,0)$ belongs to $\partial_{\mathbf{f}}[n]$. The details and the resulting automaton can be found in [52, Sec. 8.1.11]. Doing a similar job for the 2-boundary sequence, here at most 9 (and not sixteen, as in the previous example) boundary pairs may occur because $\mathbf{f}$ does not contain 11 as a factor in $\mathbf{f}$. We can check that $\partial_{\mathbf{f},2}$ is made of five different boundary sets. The resulting DFAO is depicted in Fig. 4 (the sink state reached when reading a factor 11 is not represented). We thus get the
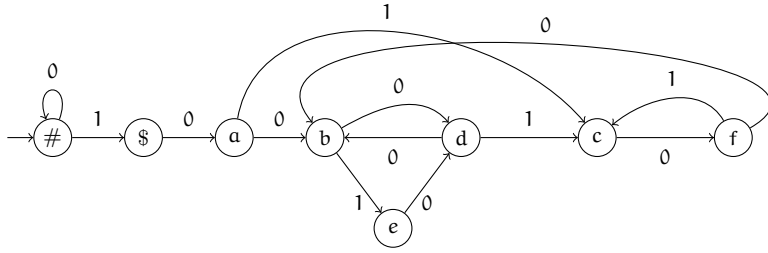
Figure 4: A DFAO producing the 2−boundary sequence of **f**.

morphism generating the 2-boundary sequence of **f** prepended with two symbols #$:

$$
\begin{cases} \# & \mapsto \#\$ \\ \$ & \mapsto a \\ a & \mapsto bc \end{cases}
\qquad
\begin{cases} b & \mapsto de \\ c & \mapsto f \end{cases}
\qquad
\begin{cases} d & \mapsto bc \\ e & \mapsto d \\ f & \mapsto bc. \end{cases}
$$

Finally, we also computed the DFAO for the boundary sequence of the Tribonacci word (reading Tribonacci expansions), the fixed point of the morphism $0 \mapsto 01$, $1 \mapsto 02$, $2 \mapsto 0$. Surprisingly, the (minimal) automaton produced by `Walnut` has 118 states. It also reveals that there are exactly seven boundary sets appearing in the boundary sequence of the Tribonacci word. One can show, using `Walnut`, that the first and second sets in the boundary sequence appear exactly once, otherwise the boundary sets appear infinitely often.

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 13) to the fact that an S-automatic word has an S-automatic boundary sequence whenever S is addable.We then show that this result does not necessarily hold for a non-addable system.

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 S-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 10 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 16. *Let* $S = (L, A, <)$ *be an addable ANS and let* $\mathbf{x}$ *be an* S*-automatic word. The boundary sequence* $\partial_{\mathbf{x}}$ *is* S*-automatic.*

Proof. Thanks to Theorem 11, the S-kernel of $\mathbf{x}$ is finite, say of cardinality $\mathfrak{m}$. Moreover, since L and $\mathcal{L}_+$ are regular, the following two sets of languages are finite by the Myhill–Nerode theorem [51, Sec. 3.9], say of cardinality k and ℓ, respectively:

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

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

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

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, $yr \in \#^*L$ and

$$\mathrm{val}_S(ws) + \mathrm{val}_S(xt) = \mathrm{val}_S(yr).$$

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 $\mathrm{reps}_S(i)$ and $\mathrm{reps}_S(i + \mathrm{val}_S(p_{s,n}s))$: we may write

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

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 $\mathrm{val}_S(p_{s,n}s)$ and is represented by a word ending with r.

For convenience, we set $L(s,t,r,n) := \mathcal{L}_+\begin{pmatrix} s \\ t \\ r \end{pmatrix}^{-1} \cap \left(\#^*p_{s,n} \times A_\#^* \times A_\#^* \right)$ for all $n \geqslant 0$. 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 $\#$ such that (w,x,y) belongs to $\mathcal{L}_+(s,t,r)^{-1}$. Similarly if y does not start with $\#$, 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 $(\#,\#,\#)$ — otherwise, different triplets may have the same numerical value. Note that $\mathrm{val}_S(x) < \mathrm{val}_S(x')$ if and only if $\mathrm{val}_S(y) < \mathrm{val}_S(y')$. Indeed, w, w' both belong to $\#^*p_{s,n}$, thus $\mathrm{val}_S(ws) = \mathrm{val}_S(p_{s,n}s) = \mathrm{val}_S(w's)$. We have

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

so,

$$\mathrm{val}_S(yr) - \mathrm{val}_S(y'r) = \mathrm{val}_S(xt) - \mathrm{val}_S(x't).$$

Since S is an ANS, if $\mathrm{val}_S(xt) > \mathrm{val}_S(x't)$, this means that, discarding the possible leading #'s because they have no effect on the evaluation, xt occurs after $x't$ in the genealogically ordered language. So x has to be genealogically larger than x'. Again discarding the possible leading #'s, the above equality means that x is genealogically less than x' if and only if the same holds for y and y'. We can thus order $L(s, t, r, n)$ by listing in increasing genealogical order the second component of the elements, 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 $j \geqslant 0$. After removing the leading #'s, the word x_j belongs to $Lt^{-1} \cup \{\varepsilon\}$, which can also be genealogically ordered. 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

$$\mathcal{L}_+(s, t, r)^{-1} = \mathcal{L}_+(s', t', r')^{-1}, \quad Ls^{-1} = Ls'^{-1}, \quad \text{and} \quad Lt^{-1} = Lt'^{-1}$$

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

$$\mathcal{L}_+(s, t, r)^{-1} = \mathcal{L}_+(s', t', r')^{-1}, \quad Ls^{-1} = Ls'^{-1}, \quad \text{and} \quad Lr^{-1} = Lr'^{-1}$$

then, for all n, the maps $\mu_{s,t,r,n}$ and $\mu_{s',t',r',n}$ are the same.

We now obtain

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

$$= \bigcup_{(t,r)\in\mathcal{P}(s)} \left\{ (\mathbf{x}[\mathrm{val}_S(x_j t)], \mathbf{x}[\mathrm{val}_S(y_j r)]) \;\middle|\; \begin{pmatrix} w_j \\ x_j \\ y_j \end{pmatrix} \in L(s, t, r, n), j \geqslant 0 \right\}$$

$$= \bigcup_{(t,r)\in\mathcal{P}(s)} \{((\mathbf{x} \circ t)[\lambda_{s,t,r,n}(j)], (\mathbf{x} \circ r)[\mu_{s,t,r,n}(j)]) \mid j \geqslant 0\}.$$

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

$$\mathcal{L}_+\begin{pmatrix} s \\ t \\ r \end{pmatrix}^{-1} = \mathcal{L}_+\begin{pmatrix} s' \\ t' \\ r' \end{pmatrix}^{-1}, \quad Ls^{-1} = Ls'^{-1}, \quad Lt^{-1} = Lt'^{-1}, \quad Lr^{-1} = Lr'^{-1}, \tag{2}$$

$$\mathbf{x} \circ t = \mathbf{x} \circ t', \quad \text{and} \quad \mathbf{x} \circ r = \mathbf{x} \circ r'.$$

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

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

Now if s and s' are such that $\mathsf{L}s^{-1} = \mathsf{L}s'^{-1}$ 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 $\mathsf{k} \cdot 2^{\ell \mathsf{k}^3 \mathsf{m}^2}$). $\qquad\square$

3.2 A family of non-addable systems

In this section, we show that the addability assumption on the numeration system is not necessary for the boundary sequence of an automatic word to be itself automatic. With Examples 18 and 22, we consider S-automatic sequences based on a non-addable ANS S but such that the corresponding boundary sequences are still S-automatic. The first lemma is merely an observation that we will frequently use.

Lemma 17. *Let $\mathbf{w} \in \{0,1\}^{\mathbb{N}}$ be an aperiodic binary word having arbitrarily long blocks of $0s$. Its boundary sequence $\partial_{\mathbf{w}}$ is over the alphabet $\{\mathfrak{a}, \mathfrak{b}\}$ where $\mathfrak{a} := \{(0,0),(0,1),(1,0)\}$ and $\mathfrak{b} := \{0,1\} \times \{0,1\}$. We have $\partial_{\mathbf{w}}[\mathsf{k}] = \mathfrak{b}$ if and only if there exists $\mathsf{m} > \mathsf{n} \geqslant 0$ such $\mathbf{w}[\mathsf{m}] = \mathbf{w}[\mathsf{n}] = 1$ and $\mathsf{k} = \mathsf{m} - \mathsf{n}$.*

Proof. By assumption, $\mathbf{w}$ contains factors of the form $0^{\mathsf{k}+1}$, $0^{\mathsf{k}}1$ and 10^{k} for all $\mathsf{k} \geqslant 1$. So a boundary set can either be $\mathfrak{a}$ or $\mathfrak{b}$. In $\mathbf{w}$, a window of length k will start with 1 and is followed by a 1 only if there exists n such that $\mathbf{w}[\mathsf{n}] = 1$ and $\mathbf{w}[\mathsf{n} + \mathsf{k}] = 1$. $\qquad\square$

Let s be an integer. In the next three examples, we consider morphic words $\mathbf{w}_s$ from the same family. They are the image under the same coding (up to erasing the first symbol) of a fixed point of $\mathsf{g}_s \colon 0 \mapsto 01, 1 \mapsto 12^s, 2 \mapsto 2$, for $s \geqslant 1$. We show that the corresponding boundary sequences $\partial_{\mathbf{w}_s}$ may exhibit quite different behaviors: for $s = 1$, it is constant; for $s = 2$, it is periodic of period 4, and for $s \geqslant 3$, it is aperiodic.

Example 18. Consider the morphisms $\mathsf{g}_1 \colon 0 \mapsto 01, 1 \mapsto 12, 2 \mapsto 2$ and $\mathsf{f} \colon 0 \mapsto \varepsilon, 1 \mapsto 1, 2 \mapsto 0$, and the word

$$\mathbf{w}_1 = \mathsf{f}(\mathsf{g}_1^{\omega}(0)) = 11010010001000010000\cdots .$$

It is the characteristic sequence of triangular numbers ([53, A000217]). A *triangular number* is any integer of the form $\mathsf{T}_{\mathsf{n}} := \binom{\mathsf{n}+1}{2} = \frac{\mathsf{n}(\mathsf{n}+1)}{2}$ for $\mathsf{n} \geqslant 0$. The sequence $(\mathsf{T}_{\mathsf{n}})_{\mathsf{n} \geqslant 0}$ starts with $0, 1, 3, 6, 10, 15, 21, 28, 36, 45, 55$.

The ANS $\mathsf{S} = (\alpha^* \beta^*, \{\alpha, \beta\}, \alpha < \beta)$ is known to be non-addable, [28, Thm. 17]. The reason is that multiplication by a constant generally does not preserve S-recognizability, hence addition cannot have this property.

Proposition 19. *Let $\mathsf{S} = (\alpha^* \beta^*, \{\alpha, \beta\}, \alpha < \beta)$. The characteristic sequence $\mathbf{w}_1$ of the set of triangular numbers given in Example 18 is S-automatic. The boundary sequence $\partial_{\mathbf{w}_1}$ is S-automatic. In particular, it is constant.*

Proof. Let s be a suffix of a word in $\alpha^* \beta^*$. We make use of the same notation as in the proof of Theorem 16. Let $\mathsf{p}_{s,\mathsf{n}}$ be the nth word in $\alpha^* \beta^* s^{-1}$, for $\mathsf{n} \geqslant 0$. As in Eq. (1), the nth term of an element of the S-kernel of $\partial_{\mathbf{w}_1}$ is given by

$$(\partial_{\mathbf{w}_1} \circ s)[\mathsf{n}] = \{\mathbf{w}_1[i]\mathbf{w}_1[i + \mathrm{val}_{\mathsf{S}}(\mathsf{p}_{s,\mathsf{n}}s)] \mid i \geqslant 0\}. \tag{3}$$

For the numeration language of interest, the admissible suffixes s are of the form $\alpha^\ell \beta^k$ for some $\ell, k \geqslant 0$. If $\ell > 0$, then $p_{s,n} = \alpha^n$ and

$$(\partial_{\mathbf{w}_1} \circ \alpha^\ell \beta^k)[n] = \{\mathbf{w}_1[i] \mathbf{w}_1[i + T_{n+\ell+k} + k] \mid i \geqslant 0\} \qquad (4)$$

because

$$\mathrm{val}_S(\alpha^i \beta^j) = \frac{1}{2}(i+j)(i+j+1) + j = T_{i+j} + j$$

for all $i, j \geqslant 0$ (see [45, Ex. 2.18]). By Lemma 17, $(\partial_{\mathbf{w}_1} \circ \alpha^\ell \beta^k)[n]$ always contains $(0,0)$, $(0,1)$, $(1,0)$. For all $n \geqslant 0$, there exists j such that $T_{n+\ell+k} + k = T_{j+1} - T_j = j + 1$. Taking $i = T_j$ in Eq. (4) shows that $(1,1)$ also belongs to $(\partial_{\mathbf{w}_1} \circ \alpha^\ell \beta^k)[n]$ which is thus equal to the set b. So the sequence $\partial_{\mathbf{w}_1} \circ \alpha^\ell \beta^k$ is constant.

Now, consider a suffix s of the form β^k, then $p_{s,n} = \alpha^i \beta^j$ for which $\mathrm{val}_S(p_{s,n}) = n$. We have, for $i, j \geqslant 0$,

$$(\partial_{\mathbf{w}_1} \circ \beta^k)[\mathrm{val}_S(\alpha^i \beta^j)] = \big\{\mathbf{w}_1[t] \mathbf{w}_1[t + \mathrm{val}_S(\alpha^i \beta^{j+k})] \mid t \geqslant 0\big\}$$

and again $(1,1)$ belongs to this set for a convenient choice of t. As a conclusion, the S-kernel contains a unique constant sequence b^ω so the boundary sequence is S-automatic. In particular, we have shown that $\partial_{\mathbf{w}}$ is constant (for the choice of suffix $s = \varepsilon$). $\qquad \square$

For the word $\mathbf{w}_1$, we can go further and prove the S-automaticity of its ℓ-boundary sequence.

Proposition 20. *Let* $S = (\alpha^* \beta^*, \{\alpha, \beta\}, \alpha < \beta)$ *and let* $\ell \geqslant 2$. *The* ℓ*-boundary sequence* $\partial_{\mathbf{w}_1, \ell}$ *of the characteristic sequence* $\mathbf{w}_1$ *of the set of triangular numbers given in Example 18 is* S*-automatic.*

Proof. Since $\mathbf{w}_1$ is the characteristic sequence of the triangular numbers, its prefix of length $T_n + 1$ ($n \geqslant 0$) ends with 1 and contains $n + 1$ occurrences of 1; more precisely we have

$$\mathbf{w}_1 = \prod_{n \geqslant 0} (10^n). \qquad (5)$$

Now consider the ℓ-boundary sequence $\partial_{\mathbf{w}_1, \ell}$ of $\mathbf{w}_1$. Let $n \geqslant \ell$ and consider a length-ℓ factor u of $\mathbf{w}_1$. Assume first that u contains at most one letter 1. Then u can take two forms.

- If $u = 0^\ell$, then the pairs $(u, 0^\ell)$, $(u, 0^i 10^{\ell-i-1})$ for $i \in \{0, 1, \ldots, \ell-1\}$ all belong to $\partial_{\mathbf{w}_1, \ell}[n]$.

- If $u = 0^i 10^{\ell-i-1}$ for some $i \in \{0, 1, \ldots, \ell-1\}$, then Eq. (5) implies that the pairs $(u, 0^\ell)$, $(u, 0^j 10^{\ell-j-1})$ for $j \in \{0, 1, \ldots, \ell-1\}$ all belong to $\partial_{\mathbf{w}_1, \ell}[n]$ if $n \geqslant 2\ell$. Indeed, let $0 \leqslant i, j < \ell$. With $0^i 10^{\ell-i-1} v 0^j 10^{\ell-j-1}$ a factor of length $n + \ell$, we must have $|v| \geqslant 2(i+1) - \ell - j$ and such a factor v clearly exists when this condition is satisfied. Taking $i = \ell - 1$ and $j = 0$ gives the maximum length requirement $|v| \geqslant \ell$, so the claim follows for $n \geqslant 2\ell$.

Therefore, the length-ℓ factors containing at most one letter 1 have the same contribution towards every boundary set in $\partial_{\mathbf{w}_1,\ell}[n]$ for $n \geqslant 2\ell$. In other words, since $\mathbf{w}_1$ has length-ℓ factors containing at least two letters 1, two boundary sets $\partial_{\mathbf{w}_1,\ell}[n]$ and $\partial_{\mathbf{w}_1,\ell}[p]$ may differ on the length-ℓ factors containing at least two letters 1.

We now examine the contribution to $\partial_{\mathbf{w}_1,\ell}[n]$ of a length-ℓ factor u containing at least two occurrences of 1. Due to Eq. (5) again, u only appears once in $\mathbf{w}_1$, so there is a unique factor v of $\mathbf{w}_1$ such that the pair (u, v) belongs to $\partial_{\mathbf{w}_1,\ell}[n]$. From Eq. (5), one sees that long stretches of letters 0 appear in $\mathbf{w}_1$. Notice that $T_{\ell-2}$ is the last position (starting at 0) of a length-ℓ factor containing at least two occurrences of 1 in $\mathbf{w}_1$ (this is the factor $10^{\ell-2}1$). Then $T_{\ell-2} + \ell = T_{\ell-1} + 1$. We compute $\partial_{\mathbf{w}_1,\ell}[n]$ for n ranging into two intervals: either n belongs to

$$I_1 := [T_m + 1, T_{m+1} - T_{\ell-1} - 1]$$

or

$$I_2 := [T_{m+1} - T_{\ell-1}, T_{m+1}]$$

for some $m > T_{\ell-1}$. (For I_1 to be non-empty, we must have $T_{m+1} - T_{\ell-1} - 1 - T_m - 1 = m - T_{\ell-1} - 1 \geqslant 0$.)

First interval. Let u be a factor of $\mathbf{w}_1$ such that $|u|_1 \geqslant 2$, and let uyv be the unique factor of length $n + \ell$, with $|y| = n$, starting with u. We claim that $v = 0^\ell$. Notice that the first letter of this particular occurrence of v appears at a position in the interval

$$[T_m + 1, T_{\ell-2} + T_{m+1} - T_{\ell-1} - 1] = [T_m + 1, T_{m+1} - \ell].$$

Hence $v = 0^\ell$. (See Example 21 for an illustration.)

Second interval. Let $n = T_{m+1} - i$ and $n' = T_{m+2} - i$ for some m and $i \in \{0, 1, \ldots, T_{\ell-1} - 1\}$. We claim that $\partial_{\mathbf{w}_1,\ell}[n] = \partial_{\mathbf{w}_1,\ell}[n']$. Suppose that $|u|_1 \geqslant 2$. Now let $y, y', v, v' \in \{0, 1\}^*$ be words such that $uyv, uy'v'$ are factors of $\mathbf{w}_1$ with $|y| = n - \ell$, $|y'| = n' - \ell$ and $|v| = \ell = |v'|$. Since $|uy| = n = T_{m+1} - i$ and $|uy'| = n' = T_{m+2} - i$, we have $v = v'$ (note that the sequence of first difference $(T_{n+1} - T_n)_{n \geqslant 0}$ goes through all positive integers). Therefore, since the length of I_2 is constant (and equal to $T_{\ell-1} + 1$), there exists a word w of length $T_{\ell-1} + 1$ such that $\prod_{n \in I_2} \partial_{\mathbf{w}_1,\ell}[n] = w$.

All in all, we have shown that $\partial_{\mathbf{w}_1,\ell}$ is of the form $p \prod_{n \geqslant 1} c^n w$ for some word p, a letter c, and a word w of length $T_{\ell-1} + 1$. It follows that $\partial_{\mathbf{w}_1,\ell}$ is S-automatic. Indeed, we have $\{\mathrm{rep}_S(T_m) \mid m \geqslant 0\} = \alpha^*$ and $\{\mathrm{rep}_S(T_{m+1} - i) \mid m > T_{\ell-1}\} = \alpha^i \beta^{\ell-i} \beta^*$ for $i > 0$. Since there are only finitely many values of i for which the corresponding boundary sets are distinct, $\partial_{\mathbf{w}_1,\ell}$ is S-automatic. $\square$

In the following example, we illustrate the proof of the previous result for several values of ℓ.

Example 21. First, consider $\ell = 2$. Let us define the boundary sets

$$\mathfrak{a} := \{(00,00), (00,01), (00,10), (01,00), (10,00), (10,01), (10,10), (11,01)\},$$
$$\mathfrak{b} := \{(00,00), (00,01), (00,10), (01,00), (10,00), (01,01), (10,01), (10,10), (11,10)\},$$
$$\mathfrak{c} := \{(00,00), (00,01), (00,10), (01,00), (10,00), (01,01), (01,10), (10,01), (10,10), (11,00)\},$$
$$\mathfrak{d} := \{(00,00), (00,01), (00,10), (01,00), (10,00), (01,01), (01,10), (10,01), (10,10), (11,01)\},$$
$$\mathfrak{e} := \{(00,00), (00,01), (00,10), (01,00), (10,00), (01,01), (01,10), (10,01), (10,10), (11,10)\}.$$

In this case, the last occurrence of a factor $\mathfrak{u}$ of length 2 with $|\mathfrak{u}|_1 = 2$ appears at position $\mathsf{T}_{\ell-2} = 0$ since $\mathbf{w}_1 = 11010010001\cdots$. The boundary set corresponding to I_1 is $\mathfrak{c}$ and those corresponding to I_2 are $\mathfrak{d}$ and $\mathfrak{e}$. Note that these sets are all distinct, and they differ precisely on the set of the form $(11, v)$. For $\mathfrak{m} = 2 = \mathsf{T}_1 + 1$, the two intervals become $\mathsf{I}_1 = [4,4]$ and $\mathsf{I}_2 = [5,6]$, and for $\mathfrak{m} = 3$, we obtain $\mathsf{I}_1 = [7,8]$ and $\mathsf{I}_2 = [9,10]$. We have $\partial_{\mathbf{w}_1,2} = \mathfrak{a}\mathfrak{b} \prod_{n \geqslant 1} (\mathfrak{c}^n \mathfrak{d}\mathfrak{e})$, for which $\mathfrak{p} = \mathfrak{a}\mathfrak{b}$, $\mathfrak{c} = \mathfrak{c}$, and $w = \mathfrak{d}\mathfrak{e}$.

Similarly, for $\ell = 3$, we have the last occurrence of a length-ℓ factor $\mathfrak{u}$ with $|\mathfrak{u}|_1 \geqslant 2$ appears at position $\mathsf{T}_{\ell-2} = 1$. For $\mathfrak{m} = \mathsf{T}_2 + 1 = 4$, we obtain $\mathsf{I}_1 = [11,11]$ and $\mathsf{I}_2 = [12,15]$. Looking at the first few letters of $\mathbf{w}_1$, one can obtain the following pairs belonging to $\partial_{\mathbf{w}_1,3}[n]$ for $n \in [11,15]$:

$(\mathfrak{u}, v)$	$n \in [11,15]$ such that $(\mathfrak{u},v) \in \partial_{\mathbf{w}_1,3}[n]$
$(110, 000)$	$11, 12$
$(110, 001)$	13
$(110, 010)$	14
$(110, 100)$	15
$(101, 000)$	$11, 15$
$(101, 001)$	12
$(101, 010)$	13
$(101, 100)$	14

So, for instance $\partial_{\mathbf{w}_1,3}[11]$ and $\partial_{\mathbf{w}_1,3}[12]$ agree on $(110,000)$ but not on $(101,000)$. Define $\mathfrak{c} = \mathfrak{t} \cup \{(101,000),(110,000)\}$ and

$$w_1 = \mathfrak{t} \cup \{(101,001),(110,000)\}, \quad w_2 = \mathfrak{t} \cup \{(101,010),(110,001)\},$$
$$w_3 = \mathfrak{t} \cup \{(101,100),(110,010)\}, \quad w_4 = \mathfrak{t} \cup \{(101,000),(110,100)\},$$

where

$$\mathfrak{t} = \{(0^3,0^3),(0^3,0^21),(0^3,010),(0^3,10^2),(0^21,0^3),(0^21,0^21),(0^21,010),(0^21,10^2),$$
$$(010,0^3),(010,0^21),(010,010),(010,10^2),(10^2,0^3),(10^2,0^21),(10^2,010),(10^2,10^2).\}$$

From the previous table, we have $\partial_{\mathbf{w}_1,3}[11;15] = \mathfrak{c}w_1w_2w_3w_4$.

Finally, we consider the case $\ell = 5$ for which $\mathsf{T}_3 = 6$. For $\mathfrak{m} = \mathsf{T}_4 + 1 = 11$, we obtain $\mathsf{I}_1 = [67,67]$ and $\mathsf{I}_2 = [68,78]$. In the following table are displayed the pairs $(\mathfrak{u}_j, v)$

belonging to $\partial_{\mathbf{w}_1,5}[n]$:

$n \in I_1 \cup I_2$	$u_0 = 11010$ $p_0 = 0$	$u_1 = 10100$ $p_1 = 1$	$u_2 = 01001$ $p_2 = 2$	$u_3 = 10010$ $p_3 = 3$	$u_4 = 10001$ $p_4 = 6$
67	0^5	0^5	0^5	0^5	0^5
68	0^5	0^5	0^5	0^5	0^41
69	0^5	0^5	0^5	0^5	0^310
70	0^5	0^5	0^5	0^5	0^210^2
71	0^5	0^5	0^5	0^41	010^3
72	0^5	0^5	0^41	0^310	10^4
73	0^5	0^41	0^310	0^210^2	0^5
74	0^41	0^310	0^210^2	010^3	0^5
75	0^310	0^210^2	010^3	10^4	0^5
76	0^210^2	010^3	10^4	0^5	0^5
77	010^3	10^4	0^5	0^5	0^5
78	10^4	0^5	0^5	0^5	0^5

For instance, we see that the sets $\partial_{\mathbf{w}_1,5}[n]$ for $n \in [74, 78]$ differ on the pair (u_0, v). However $\partial_{\mathbf{w}_1,5}[n]$ for $n \in [67, 73]$ all contain $(u_0, 0^5)$ so u_0 cannot tell them apart. To that aim, one has to go through all columns in the previous table, therefore covering all possible values of u_i.

Example 22. Let $S = (\alpha^*\beta^* \cup \beta^*\gamma^*, \{\alpha, \beta, \gamma\}, \alpha < \beta < \gamma)$ be an abstract numeration system whose language has exactly $2n + 1$ words of length n. For a construction of regular languages with a specific polynomial growth, see [44]. Consider the S-automatic word given by the characteristic sequences of the words from the sublanguage α^* within $\alpha^*\beta^* \cup \beta^*\gamma^*$: $\mathbf{w}_2 = 1100100001000000 \cdots$. This is exactly the characteristic sequence of the set of squares. This word is also obtained using the morphisms $g_2 \colon 0 \mapsto 01, 1 \mapsto 122, 2 \mapsto 2$ and $f \colon 0 \mapsto \varepsilon, 1 \mapsto 1, 2 \mapsto 0$, $\mathbf{w}_2 = f(g_2^\omega(0))$. Notice that again the ANS S is non-addable, this follows from [43, Thm. 15].

Proposition 23. *Let* $S = (\alpha^*\beta^* \cup \beta^*\gamma^*, \{\alpha, \beta, \gamma\}, \alpha < \beta < \gamma)$. *The boundary sequence* $\partial_{\mathbf{w}_2}$ *of the characteristic sequence of the set of squares given in Example 22 is S-automatic. In particular, it is periodic with period* `babb`.

We provide two proofs, the first one is generic. It aims to show the finiteness of the S-kernel of $\partial_{\mathbf{w}_2}$ without explicitly determining the boundary sequence. The second one is less systematic but directly shows periodicity.

Proof sketch. To prove that the S-kernel is finite, we first guess that it contains 14 elements. We have computed prefixes of elements of the S-kernel with different suffixes given in Table 2. For the system S, since values are given by positions within the genealogically ordered language, we easily get

$$\mathrm{val}_S(\alpha^i\beta^j) = (i + j)^2 + j \quad \text{and} \quad \mathrm{val}_S(\beta^j\gamma^k) = (j + k)^2 + j + 2k.$$

ε	b	a	b	b	b	a	b	b	b	a	b	b	b	a	b	b	b	a	b
α	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b
β	a	b	a	a	b	b	b	a	b	b	a	b	b	b	a	b	a	b	b
β^2	a	b	b	a	b	b	b	b	b	a	a	b	b	b	a	b	b	b	a
β^3	b	b	b	b	b	b	a	b	b	b	a	b	b	a	b	b	b	b	b
β^4	b	b	a	b	b	a	b	a	b	b	b	b	a	b	b	b	a	b	b
γ	b	b	b	b	a	b	b	a	b	b	b	b	b	a	b	b	b	b	a
γ^2	b	a	b	a	b	b	b	b	a	b	b	b	a	b	b	a	b	b	b
ab	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b
$\alpha^2 b$	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a
$b\gamma^2$	a	a	b	b	a	a	b	b	a	a	b	b	a	a	b	b	a	a	b
$\beta^2\gamma^2$	a	b	b	a	a	b	b	a	a	b	b	a	a	b	b	a	a	b	b
$\beta^3\gamma^2$	b	b	a	a	b	b	a	a	b	b	a	a	b	b	a	a	b	b	a
$\beta^4\gamma^2$	b	a	a	b	b	a	a	b	b	a	a	b	b	a	a	b	b	a	a

Table 2: Prefixes of elements of the form $\partial_{\mathbf{w}_2} \circ s$.

The suffixes to consider are of the form

$$\beta^k, \ \alpha^\ell\beta^k, \ \gamma^k, \ \beta^\ell\gamma^k \ \text{ with } k, \ell \geqslant 0.$$

By Lemma 17, every boundary set contains at least $(0,0)$, $(0,1)$, $(1,0)$. The only question is therefore to determine whether $(1,1)$ belongs to some specific boundary set. In view of Table 2, we have to prove relations such as the following one (that we treat in details)

$$\partial_{\mathbf{w}_2} \circ \beta^j = \partial_{\mathbf{w}_2} \circ \beta^{j+4r}, \quad j \in \{1,2,3,4\}, r > 0.$$

Let $n \geqslant 0$. We still have Eq. (3)

$$(\partial_{\mathbf{w}_2} \circ \beta^j)[n] = \{\mathbf{w}_2[i]\mathbf{w}_2[i + \mathrm{val}_S(\alpha^k\beta^{\ell+j})] \mid i \geqslant 0\}$$

where $\alpha^k\beta^\ell$ is the nth word in $L(\beta^j)^{-1}$ with L being the associated ANS language. Since $L(\beta^j)^{-1} = L(\beta^m)^{-1} = \alpha^*\beta^*$ for any $j, m > 0$, we also have

$$(\partial_{\mathbf{w}_2} \circ \beta^{j+4r})[n] = \{\mathbf{w}_2[i]\mathbf{w}_2[i + \mathrm{val}_S(\alpha^k\beta^{\ell+j+4r})] \mid i \geqslant 0\}.$$

By definition of the word $\mathbf{w}_2$ and Lemma 17, $(1,1)$ belongs to the above set if and only if $\mathrm{val}_S(\alpha^k\beta^{\ell+j+4r})$ is the difference of two squares. Modulo 4, a square is congruent to either 0 or, 1. Such a difference is not congruent to 2 modulo 4. It is straightforward to express any number belonging to the other three congruence classes as the difference of two squares: $4m + 1 = (2m+1)^2 - (2m)^2$, $4m - 1 = (2m)^2 - (2m-1)^2$ and $4m = (m+1)^2 - (m-1)^2$. Observe that

$$\mathrm{val}_S(\alpha^k\beta^{\ell+j}) = (k+\ell+j)^2 + \ell + j$$

and

$$\mathrm{val}_S(\alpha^k\beta^{\ell+j+4r}) = (k+\ell+j+4r)^2 + \ell + j + 4r$$

are congruent (mod 4). So $(1,1)$ belongs to the set $(\partial_{\mathbf{w}_2} \circ \beta^j)[n]$ if and only if it belongs to $(\partial_{\mathbf{w}_2} \circ \beta^{j+4r})[n]$, leading to the conclusion.

The first few words in $L\beta^{-1}$ are $\varepsilon, \alpha, \beta$. To distinguish, as an example, the elements $\partial_{\mathbf{w}_2} \circ \beta$ and $\partial_{\mathbf{w}_2} \circ \beta^2$ of the S-kernel, it is enough to look at

$$(\partial_{\mathbf{w}_2} \circ \beta)[2] = \{\mathbf{w}_2[i]\mathbf{w}_2[i + \text{val}_S(\beta^2)] \mid i \geqslant 0\}$$

and

$$(\partial_{\mathbf{w}_2} \circ \beta^2)[2] = \{\mathbf{w}_2[i]\mathbf{w}_2[i + \text{val}_S(\beta^3)] \mid i \geqslant 0\}$$

because $\text{val}_S(\beta^2) \equiv 2 \pmod 4$ and $\text{val}_S(\beta^3) \equiv 0 \pmod 4$. So the first one is a and the second one is b (as shown in Table 2).

Proving the finiteness of the S-kernel amounts to prove relations such as:

$$\partial_{\mathbf{w}_2} \circ \alpha\beta = \partial_{\mathbf{w}_2} \circ \alpha^{2m+1}\beta^{\{1,2\}+4n}, \ \partial_{\mathbf{w}_2} \circ \alpha^2\beta = \partial_{\mathbf{w}_2} \circ \alpha^{2m}\beta^{\{1,2\}+4n}, \dots. \qquad \square$$

Here is a shorter proof because, in our particular example, the boundary sequence is periodic.

Proof. Let us show that $\partial_{\mathbf{w}_2} = (babb)^\omega$. We make use of Lemma 17: $\partial_{\mathbf{w}_2}[k] = b$ if and only if k can be written as the difference of two squares $m^2 - n^2$ with $m > n \geqslant 0$. With the same argument as in the previous proof, this holds if and only if k is not congruent to 2 modulo 4. Automaticity follows from the fact that any ultimately periodic set is S-recognizable for all ANS [28, Thm. 4]. $\qquad \square$

Remark 24. With Examples 18 and 22, we have exhibited sequences that are S-automatic for some non-addable numeration system S. One can naturally wonder if these sequences could also be T-automatic for another numeration system T being addable. Since the considered numeration systems have a polynomial growth, a Cobham-like result implies that if $\mathbf{w}_1$ (resp., $\mathbf{w}_2$) is T-automatic for some T, then T must have a polynomial growth [18, Cor. 27]. As a consequence of [43, Thm. 15], ANS with a polynomial growth are not addable. This means that Examples 18 and 22 highlight words that are S-automatic only for some non-addable numeration systems S.

As a side comment, if an addable numeration system is such that the graph of $n \mapsto T_n$ is also regular (i.e., the set of pairs $(\text{rep}(n), \text{rep}(T_n))$, where the shortest representation is conveniently padded, is a regular language), then the first order theory of $\langle \mathbb{N}, +, x^2 \rangle$ would be decidable. But this structure is equivalent to $\langle \mathbb{N}, +, \cdot \rangle$ which is well known to have an undecidable theory.

To end up this short section, we consider a third example which is a small variation of the previous one.

Example 25. For $s \geqslant 3$, take the morphic word $\mathbf{w}_s = f(g_s^\omega(0))$ where $g_s \colon 0 \mapsto 01$, $1 \mapsto 12^s$, $2 \mapsto 2$ and $f \colon 0 \mapsto \varepsilon$, $1 \mapsto 1$, $2 \mapsto 0$. For a fixed s, the word $\mathbf{w}_s$ is the characteristic word of the set of numbers of the form $P_n := \frac{n(sn-s+2)}{2}$. For example, the word $\mathbf{w}_3$ is the characteristic sequence of the set of *pentagonal numbers* ([53, A000326]), $\mathbf{w}_4$ of the *hexagonal numbers* ([53, A000384]), and $\mathbf{w}_5$ of the *heptagonal numbers* ([53, A000566]).

In the remainder of this part, we fix $s \geqslant 3$ and write $\mathbf{w} = \mathbf{w}_s$ for short. Applying Lemma 17, the boundary sequence is such that $\partial_{\mathbf{w}}[k] = b$ if and only if k can be written as

$$k = P_m - P_n = \frac{1}{2}(m - n)(s(m + n - 1) + 2) \tag{6}$$

for some integers $m > n \geqslant 0$. We say that an integer k is *representable* if there exist $m, n \in \mathbb{N}$ with $m > n$ such that the above equation holds.

Proposition 26. *The boundary sequence $\partial_{\mathbf{w}}$ is aperiodic.*

Proof. Assume first that s is odd. We make an observation about representable integers of a certain form.

Claim 27. *Let p be a prime number congruent to $1 \pmod{s}$. For any $i, j \geqslant 0$, $s^i \cdot p^j$ is representable if and only if $p^j \geqslant s\frac{s^i-1}{2} + 1$.*

Proof of claim: Notice that p is an odd prime number. Assume that $p^j \geqslant s\frac{s^i-1}{2} + 1$. Then there exists $n \geqslant 0$ such that $p^j = s\frac{s^i-1}{2} + 1 + ns$. Setting $m = n + s^i$, we find

$$
\begin{aligned}
P_m - P_n &= \frac{1}{2}(m - n)(s(m + n - 1) + 2) \\
&= \frac{1}{2}s^i(s(2n + s^i - 1) + 2) \\
&= s^i(s\frac{s^i - 1}{2} + ns + 1) \\
&= s^i p^j.
\end{aligned}
$$

Thus $s^i p^j$ is representable.

Assume then that $p^j < s\frac{s^i-1}{2} + 1$, but towards a contradiction, that $s^i p^j = P_m - P_n$ for some integers $m > n \geqslant 0$. We thus have

$$2s^i p^j = (m - n)(s(m + n - 1) + 2).$$

Notice that $s(m + n - 1) + 2 \equiv 2 \pmod{s}$. Consequently, as $s \geqslant 3$, we must have $s^i \mid m - n$. Furthermore, since $p \equiv 1 \pmod{s}$, we must have that $2 \mid s(m+n-1)+2$ due to the same observation. Therefore, we have $s(m + n - 1) + 2 = 2p^{j_1}$ and $m = n + p^{j_2}s^i$ with $j_1 + j_2 = j$. Plugging the latter into the former, we find

$$2p^{j_1} = s(2n + s^i p^{j_2} - 1) + 2 = s(s^i p^{j_2} - 1) + 2 + s2n \geqslant s(s^i - 1) + 2 > 2p^j,$$

where in the last inequality, we have used the assumption. This is a contradiction. Thus $s^i p^j$ is not representable, as claimed. ∎

Assume towards a contradiction that $\partial_{\mathbf{w}}$ is eventually periodic, i.e., $\partial_{\mathbf{w}} = uv^{\omega}$ for some finite words u, v. Let $i \geqslant 1$ be such that $s^i \geqslant |u|$. Then the previous claim and (6) imply $\partial_{\mathbf{w}}[s^i] = a$, and by assumption, $\partial_{\mathbf{w}}[s^i + n|v|] = a$ for all $n \geqslant 0$. Let however p be a prime congruent to $1 \pmod{|v|s}$ (and thus $p \equiv 1 \pmod{s}$) and $p \geqslant s\frac{s^i-1}{2} + 1$. Note

that there exist infinitely many primes of this form by Dirichlet's theorem for primes in arithmetic progressions (see, e.g., [3, Thm. 7.9]). Write $p = q \cdot |v|s + 1$. Take $n = s^{i+1}q$; then we have

$$s^i + n|v| = s^i + s^{i+1}q|v| = s^i(1 + qs|v|) = s^i p.$$

This implies that $\partial_{\mathbf{w}}[s^i + n|v|] = b$ by the above claim together with (6). This contradiction shows that $\mathbf{w}$ is aperiodic when s is odd.

Assume then that s is even, say $s = 2t$ with $t \geqslant 2$. Then we have that k is representable if and only if $k = P_n - P_m = (m - n)(t(m + n - 1) + 1)$.

Claim 28. *Let p be a prime number congruent to $1 \pmod{s}$. Let $q = 1$ if t is odd, otherwise let $q = t + 1$. Then, for all $i, j \geqslant 0$, we have that $t^i \cdot q \cdot p^j$ is representable if and only if $p^j \cdot q \geqslant t(t^i - 1) + 1$.*

Proof of claim: If $p^j \cdot q \geqslant t(t^i - 1) + 1$, then there exists $n \geqslant 0$ such that $p^j q = t(t^i - 1) + ns + 1$: indeed, if t is odd, we have $t(t^i - 1) \equiv 0 \pmod{s}$ and $p^j q = p^j \equiv 1 \pmod{s}$. If t is even, then $t(t^i - 1) \equiv t \pmod{s}$ and we have $p^j q = p^j(t + 1) \equiv t + 1 \pmod{s}$. Now set $m = n + t^i$. We thus find

$$P_m - P_n = (n - m)(t(m + n - 1) + 1) = t^i(t(2n + t^i - 1) + 1) = t^i(t(t^i - 1) + sn + 1) = t^i p^j q,$$

showing that $t^i p^j q$ is representable.

For the converse, assume again that $t^i p^j q = P_m - P_n$ but that $p^j q < t(t^i - 1) + 1$. We thus have

$$t^i p^j q = (m - n)(t(m + n - 1) + 1).$$

By inspection modulo t, we must have that $m - n = t^i p^{j_1} q_1$ and $t(m + n - 1) + 1 = p^{j_2} q_2$, where $j_1 + j_2 = j$ and $q_1 q_2 = q$. We plug in $m = t^i p^{j_1} q_1 + n$ into the second term to obtain

$$p^{j_2} q_2 = t(2n + t^i p^{j_1} q_1 - 1) + 1 = 2nt + t(t^i p^{j_1} q_1 - 1) + 1 \geqslant t(t^i - 1) + 1 > p^j q,$$

where the last inequality is obtained by using the assumption. This is a contradiction. Therefore $t^i p^j q$ is not representable, as was claimed. ∎

To conclude the proof of the proposition, assume again towards a contradiction that $\partial_{\mathbf{w}} = uv^\omega$. Let $q = 1$ if t is odd, and otherwise let $q = t + 1$. Let $i \geqslant 1$ be such that $t(t^i - 1) + 1 > q$ and $t^i \geqslant |u|$. Then by the above claim $t^i q$ is not representable. In fact, by periodicity, we have that $t^i q + n|v|$ is not representable for all $n \geqslant 0$. Let however p be a prime with $p \equiv 1 \pmod{s|v|}$ (thus $p \equiv 1 \pmod{s}$), and such that $pq \geqslant t(t^i - 1) + 1$ (again Dirichlet's theorem implies the existence of such a prime). Write $p = r \cdot s|v| + 1$ and let $n = t^i qsr$. We then have $t^i q + n|v| = t^i q + t^i qrs|v| = t^i q(1 + rs|v|) = t^i qp$, which is a representable number by the above claim. This contradiction shows that $\partial_{\mathbf{w}}$ is aperiodic. □

3.3 Non-addable systems: counterexamples

Our aim is to show that the boundary sequence of a U-automatic word is not always U-automatic. Here, we have special instances of abstract numeration systems which are, in particular, positional. So we refer to the sequence U defining the system. We give two such examples. The numeration system defined first is a variant of the base-2 system.

Example 29. Take the numeration system $(\mathsf{U}_n)_{n \geqslant 0}$ defined by $\mathsf{U}_n = 2^{n+1} - 1$ for all $n \geqslant 0$. We have $0^* \operatorname{rep}_{\mathsf{U}}(\mathbb{N}) = (0+1)^*(\varepsilon + 20^*)$. Consider the characteristic word $\mathbf{u}$ of U, i.e., $\mathbf{u}[n] = 1$ if and only if $n \in \{\mathsf{U}_j \mid j \geqslant 0\}$. The boundary sequence $\partial_{\mathbf{u}}$ starts with

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

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

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

Proposition 30. *Let* $\mathsf{U} = (2^{n+1} - 1)_{n \geqslant 0}$. *The word* $\mathbf{u}$ *from Example 29 is* U-*automatic but its boundary sequence* $\partial_{\mathbf{u}}$ *is not* U-*automatic.*

Proof. The word $\mathbf{u}$ is trivially U-automatic. By Lemma 17, we have $\partial_{\mathbf{u}}[k] = b$ if and only if k is of the form $\mathsf{U}_m - \mathsf{U}_n = 2^{m+1} - 2^{n+1}$ for some $m > n \geqslant 0$. Therefore $\partial_{\mathbf{u}}$ is U-automatic if and only if the set $X := \{\mathsf{U}_{m+r} - \mathsf{U}_m \mid m \geqslant 0, r > 0\}$ is U-recognizable (i.e., $\operatorname{rep}_{\mathsf{U}}(X)$ is regular). Set $R := \operatorname{rep}_{\mathsf{U}}(X)$ and

$$R_1 := 20^*, \quad R_2 := \bigcup_{k \geqslant 1} 1^k 0^* \operatorname{rep}_{\mathsf{U}}(k), \quad R_3 := \bigcup_{k \geqslant 1} 1^{\mathsf{U}_k - 1} \operatorname{rep}_{\mathsf{U}}(\mathsf{U}_k - 1) 0^* = \bigcup_{k \geqslant 1} 1^{\mathsf{U}_k - 1} 20^{k-1} 0^*.$$

We have $R = R_1 \cup R_2 \cup R_3$ by the following three observations. The U-representations of the elements in X for $m = 0$ and $r > 0$ are given by the words in R_1 because, in that case,

$$\operatorname{val}_{\mathsf{U}}(20^{r-1}) = \mathsf{U}_r - 1 = \mathsf{U}_r - \mathsf{U}_0.$$

For $m > 0$ and $r < \mathsf{U}_m$, i.e., $|\operatorname{rep}_{\mathsf{U}}(r)| \leqslant m - 1$, the U-representations of the elements in X are given by the words in R_2 because

$$\operatorname{val}_{\mathsf{U}}(1^r 0^{m - |\operatorname{rep}_{\mathsf{U}}(r)| - 1} \operatorname{rep}_{\mathsf{U}}(r)) = \mathsf{U}_{m+r} - \mathsf{U}_m.$$

Finally, the case $m > 0$ and $r \geqslant \mathsf{U}_m$ is handled by the words in R_3 since

$$\operatorname{val}_{\mathsf{U}}(1^{\mathsf{U}_m - 1} 20^{m + \ell - 1}) = \mathsf{U}_{m + \mathsf{U}_m + \ell} - \mathsf{U}_m.$$

An application of the pumping lemma shows that R is not regular. By contradiction, if R is regular, then $R \cap 1^* 20^*$ is regular and accepted by a DFA with t states. We conclude that there exist infinitely many integers $n_0 < n_1 < n_2 < \cdots$ and a constant C such that $1^{n_i} 20^C$ belongs to $R \cap 1^* 20^*$. This contradicts the form of the words in $R_2 \cup R_3$. Consequently, $\partial_{\mathbf{u}}$ is not U-automatic. $\qquad\square$

As a consequence of the previous proposition and Theorem 16, U is non-addable.

Remark 31. One may notice that both $\mathbf{u}$ and $\partial_{\mathbf{u}}$ are 2-automatic: this follows by the Büchi–Bruyère theorem [9] from the set

$$X := \{U_{m+r} - U_m \mid m \geqslant 0, r > 0\} = \{n \in \mathbb{N}: \partial_{\mathbf{u}}[n] = b\}$$

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 29 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 31. To this end, consider the non-addable numeration system from [22, Ex. 3] or [33, Ex. 2] defined by

$$V_0 = 1,\ V_1 = 4,\ V_2 = 15,\ V_3 = 54 \quad \text{and} \quad V_n = 3V_{n-1} + 2V_{n-2} + 3V_{n-4}, \quad \forall n \geqslant 4. \quad (7)$$

Example 32. Consider the characteristic word $\mathbf{v}$ of V, i.e., $\mathbf{v}[n] = 1$ if and only if $n \in \{V_j \mid j \geqslant 0\}$. This word is trivially V-automatic. The boundary sequence $\partial_{\mathbf{v}}$ starts with

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

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

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

Proposition 33. *Let V be the numeration system given by* (7). *The word $\mathbf{v}$ from Example 32 is V-automatic but its boundary sequence $\partial_{\mathbf{v}}$ is not V-automatic.*

Before diving into the proof, we set the stage with some remarks of the numeration system given in (7). We assume that the reader has some knowledge about β-numeration systems, see, for instance [45].

The characteristic polynomial of (7) has two real roots β and γ and two complex roots with modulus less than 1. We have $\beta \simeq 3.61645$ and $\gamma \simeq -1.09685$. The number β is neither a Pisot number nor a Salem number. It is however a Parry number, as it is readily checked that $d_\beta(1) = 3203$, where for any real number $x \in [0,1]$, we let $d_\beta(x) = c_0 c_1 \cdots$ denote the (greedy) β-expansion of x satisfying $x = \sum_{i=0}^{\infty} c_i \beta^{-i-1}$ and $x - \sum_{i=0}^{j} c_i \beta^{-i-1} < \beta^{-j-1}$ for all $j \geqslant 0$. The quasi-greedy expansion $d_\beta^*(1)$ of 1, defined as $\lim_{x \to 1^-} d_\beta(x)$, is then $(3202)^\omega$. Thus V is a Parry numeration system such that $\mathrm{rep}_V(\mathbb{N})$ is regular. In our setting, every element in $\mathbb{Q}(\beta)$ is a polynomial of degree at most 3 in $\mathbb{Q}[\beta]$.

Lemma 34 ([50, Lem. 2.2]). *Let $x \in [0,1) \cap \mathbb{Q}(\beta)$, and write $x = q^{-1} \sum_{i=0}^{3} p_i \beta^i$ for integers q and p_i. If $d_\beta(x)$ is ultimately periodic, then*

$$q^{-1} \sum_{i=0}^{3} p_i \gamma^i = \sum_{i=1}^{\infty} d_\beta(x)[i] \, \gamma^{-i}. \tag{8}$$

Proof of Proposition 33. By Lemma 17, we have $\partial_v[k] = b$ if and only if k is of the form $V_{m+r} - V_m$ for some $m \geqslant 0$ and $r > 0$. We discuss the value of r modulo 4:

$$\mathrm{rep}_V(\{V_{m+4j} - V_m \mid m \geqslant 0\}) = \{(3202)^j 0^* \colon j \geqslant 1\};$$
$$\mathrm{rep}_V(\{V_{m+4j+1} - V_m \mid m = 0,1,2\}) = \{(3202)^j(3 + 23 + 221) \colon j \geqslant 0\};$$
$$\mathrm{rep}_V(\{V_{m+4j+1} - V_m \mid m \geqslant 3\}) = \{(3202)^j(2203)0^* \colon j \geqslant 0\};$$
$$\mathrm{rep}_V(\{V_{m+4j+2} - V_m \mid m = 0,1\}) = \{(3202)^j(32 + 311) \colon j \geqslant 0\};$$
$$\mathrm{rep}_V(\{V_{m+4j+2} - V_m \mid m \geqslant 2\}) = \{(3202)^j(3103)0^* \colon j \geqslant 0\}.$$

The first equality comes from the fact that $(3202)^j 0^m$ is a greedy representation and

$$\mathrm{val}_V((3202)^j 0^m) + V_m = \mathrm{val}_V((3202)^{j-1}(3203)0^m) = V_{m+4j}.$$

The reasoning is similar for the third and fifth equalities. For the third, we get

$$\mathrm{val}_V((3202)^j(2203)0^m) + V_{m+3} = \mathrm{val}_V((3202)^{j-1}(3203)0^m) = V_{m+4j+4},$$

which means that $\mathrm{rep}_V(V_{m+4j+4} - V_{m+3})$ is $(3202)^j(2203)0^m$ because it is lexicographically less than $d_\beta^*(1)$ and thus a valid expansion. Similarly, for the fifth, we have

$$\mathrm{val}_V((3202)^j(3103)0^m) + V_{m+2} = \mathrm{val}_V((3202)^{j-1}(3203)0^m) = V_{m+4j+4}.$$

Finally, we prove that, for all k, there exists M such that for all $m \geqslant M$ and $j \geqslant 0$, there exists a suffix $t_{m,j} \in \{0,1,2,3\}^*$ of length $m - k - 2$ such that

$$\mathrm{rep}_V(V_{m+4j+3} - V_m) = (3202)^j \, d[0] \cdots d[k-1] \, t_{m,j}, \tag{9}$$

where $d = 3131202300020211200210312213101221120211\cdots$ is the β-expansion of

$$1 - 1/\beta^3 = (18 - 7\beta - 6\beta^2 + 2\beta^3)/9.$$

Roughly speaking, $\mathrm{rep}_V(V_{m+4j+3} - V_m)$ starts with $(3202)^j$ but then, for increasing values of m, the corresponding words share longer and longer prefixes of d. See Table 3. Let us first focus on the case $j = 0$, i.e., on the V-representation of $V_{m+3} - V_m$. Let $k > 0$. Proceed by contradiction and assume that for some $t < k$, $d[0] \cdots d[t]$ is not a greedy expansion, i.e.,

$$V_{m+3} - V_m - \sum_{i=0}^{t} d[i] \, V_{m+2-i} \geqslant V_{m+2-t}.$$

m															
0	1	0	0	0											
1	3	2	0	0											
2	3	1	3	1	3										
3	3	1	3	1	2	2									
4	3	1	3	1	2	1	0								
5	3	1	3	1	2	0	3	1							
6	3	1	3	1	2	0	2	3	1						
7	3	1	3	1	2	0	2	3	0	1					
8	3	1	3	1	2	0	2	3	0	0	1				
9	3	1	3	1	2	0	2	3	0	0	1	0			
10	3	1	3	1	2	0	2	3	0	0	0	2	2		
11	3	1	3	1	2	0	2	3	0	0	0	2	1	0	
12	3	1	3	1	2	0	2	3	0	0	0	2	0	2	2
13	3	1	3	1	2	0	2	3	0	0	0	2	0	2	1	3
14	3	1	3	1	2	0	2	3	0	0	0	2	0	2	1	1	3

Table 3: The V-representations of $V_{m+3} - V_m$ for $m = 0, \ldots, 14$.

Dividing both sides by V_{m+3} and letting m tend to infinity, we get

$$1 - 1/\beta^3 - \sum_{i=0}^{t} \mathbf{d}[i]/\beta^{i+1} \geqslant 1/\beta^{t+1},$$

contradicting the fact that $\mathbf{d}$ is the β-expansion of $1 - 1/\beta^3$. Now, for $j \geqslant 1$, write

$$V_{m+4j+3} - V_m = \sum_{i=1}^{j} (V_{m+4i+3} - V_{m+4(i-1)+3}) + V_{m+3} - V_m.$$

By using the recurrence relation defining V, it is clear that

$$\mathrm{rep}_V(V_{m+4i+3} - V_{m+4(i-1)+3}) = 3202\,0^{m+4(i-1)+2}.$$

Hence $\mathrm{rep}_V(V_{m+4j+3} - V_m)$ has the expected form (9).

We now show that $\mathbf{d}$ is not ultimately periodic. We apply Lemma 34 for $x = 1 - 1/\beta^3$. The left-hand side in (8) is approximately 1.75. Since $|\gamma| > 1$, the right-hand side converges (absolutely) and the first few digits of its limit are -3.57. Hence $\mathbf{d}$ is not ultimately periodic.

To conclude the proof, we apply the pumping lemma to show that the language $R := \mathrm{rep}_V(\{V_{m+r} - V_m \mid m \geqslant 0, r > 0\})$ is not regular. Proceed by contradiction. Suppose that R is accepted by a DFA with ℓ states. Then there exist words u, v, w with $0 < |v| \leqslant \ell$ and $\mathbf{d}$ has uv as prefix such that, for all n, $uv^n w$ belongs to R. This is a contradiction because $\mathbf{d}$ is not periodic. □

Remark 35. In the above proof, it is interesting to note that the non-regularity of the language R is really associated with $V_{m+r} - V_m$ for r congruent to 3 modulo 4. Indeed, we have used the fact that $d_\beta(1-1/\beta^3)$ is not ultimately periodic whereas $d_\beta(1-1/\beta) = 2203$, $d_\beta(1-1/\beta^2) = 3103$ and $d_\beta(1-1/\beta^4) = 3202$.

Remark 36. 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 37 and Proposition 47) 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 [30, §2]. Let $\alpha, \rho \in \mathbb{T} := [0,1)$ 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 $n \geqslant 0$. 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 $n \geqslant 0$ and some $\mathcal{B} \colon A^r \to B$. Let $T \colon A^\mathbb{N} \to A^\mathbb{N}$ denote the shift map $Tx_0 x_1 x_2 \cdots = x_1 x_2 \cdots$.

Theorem 37. *For a Sturmian word* $\mathbf{s}$ *of slope* α *(and intercept* ρ*) and* $\ell \geqslant 1$*, 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 $\{\{-j\alpha\}\}_{j=0}^\ell$ 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 38. *Let* $n \geqslant \ell$ *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* $R_\alpha^n(I_u) \cap I_v \neq \emptyset$*.*

Proof of claim: We have $(u,v) \in \partial_{\mathbf{x},\ell}[n]$ if and only if there exists i such that $R_\alpha^i(\rho) \in I_u$ and $R_\alpha^{i+n}(\rho) \in I_v$, or equivalently, $R_\alpha^i(\rho) \in I_u \cap R_\alpha^{-n}(I_v)$. Notice that the intersection is a

finite union of (possibly empty) intervals. Since the set $(R_\alpha^i(\rho))_{i \in \mathbb{N}}$ is dense in $\mathbb{T}$, it follows that there exists i such that $R_\alpha^i(\rho) \in I_u \cap R_\alpha^{-n}(I_v)$ if and only if $I_u \cap R_\alpha^{-n}(I_v) \neq \emptyset$. The claim follows by applying the isomorphism R_α^n to the intersection. ∎

The endpoints of I_u are of the form i_j and i_{j+1} for some $j \in \{0, \ldots, \ell\}$. Hence, for $n \geqslant \ell$, 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 45. 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 40.

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

$$R_\alpha^n(\{-j\alpha\}) \in I_{w_{\sigma(j)}} \quad \forall j \in \{0, \ldots, \ell\}. \tag{10}$$

The realizable such configurations in (10) 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 40 (and Example 41) for an illustration of the construction.

Definition 39. Let $\sigma \in T_\ell$ be such that (10) 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 40. The Fibonacci word $\mathbf{f}$ is $\mathbf{s}_{\alpha,\alpha}$ for $\alpha = (3 - \sqrt{5})/2 \simeq 0.382$. In Fig. 5, 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 $(0, 1, 2, 3, 4) \mapsto (2, 0, 3, 1, 4)$. For any constellation inducing σ_n, we see the pairs belonging to $\partial_{\sigma_n} = \partial_{\mathbf{f},4}[17]$ from Fig. 5: 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 2, the five sets $a_1, \ldots, a_5$ correspond to the situations depicted from left to right in Fig. 6. 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.

We give an accompanying example to Example 40 for the reader to clarify the notion on constellations.
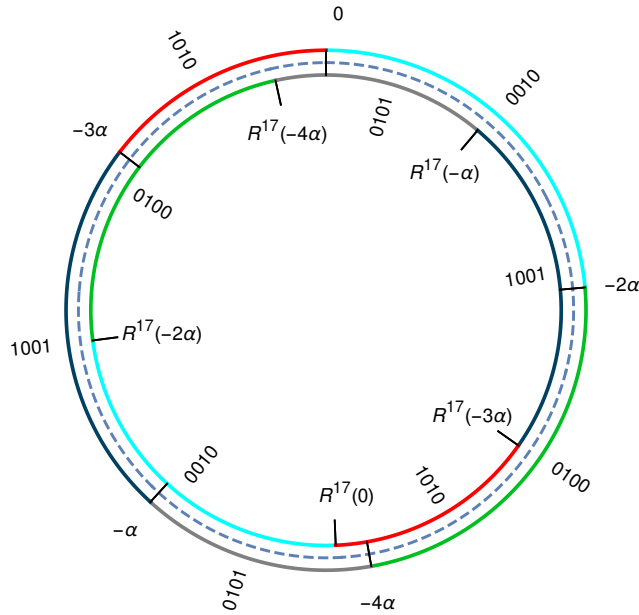
Figure 5: A constellation for $\alpha = (3 - \sqrt{5})/2$, $\ell = 4$ and $n = 17$.
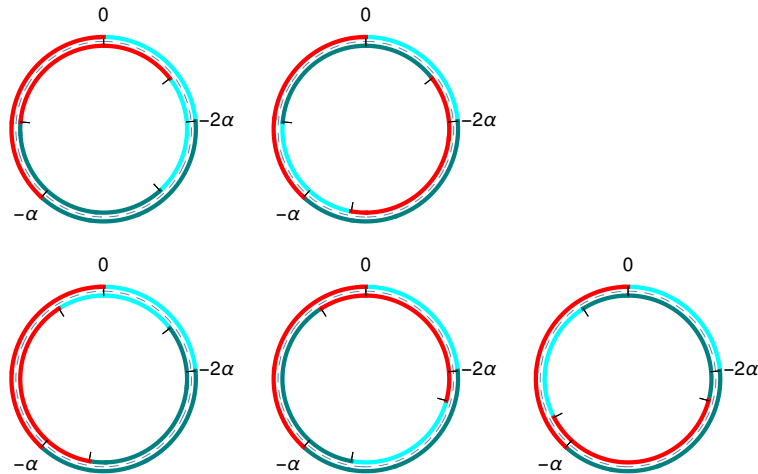

Figure 6: Some constellations for $\alpha = (3 - \sqrt{5})/2$ and $\ell = 2$ inducing the five maps σ_n sending $(0, 1, 2)$, resp., to $(0, 2, 1)$, $(1, 0, 2)$, $(2, 1, 0)$, $(1, 2, 1)$, $(2, 1, 2)$.

Example 41. What matters to determine the pairs belonging to the ℓ-boundary sequence are the non-empty intersections of the form $R_\alpha^n(I_u) \cap I_v$. There are situations where $R_\alpha^n(I_u) \subset I_v$ or $R_\alpha^n(I_u) \supset I_v$, whence σ_n is neither injective nor surjective. For instance, this is the case for the last two constellations in Fig. 6 (we have σ_n equals $(0, 1, 2) \mapsto (1, 2, 1)$, and $(0, 1, 2) \mapsto (2, 1, 2)$, respectively). With $\alpha = (\pi - 3)/2 \simeq 0.0708$ and $\ell = 5$, the partition of $\mathbb{T}$ is made of 5 short intervals of length α and one large interval of length

$1 - 5\alpha > 0.5$. In Fig. 7, we see that five or four "short" rotated intervals are included in the same large interval (for n equal to 21 and 10 respectively). In particular, counting the number of matching pairs of colors around the circle, we see that $\partial_{\mathbf{x},5}[5] = \partial_{\mathbf{x},5}[21]$ with cardinality 11 and $|\partial_{\mathbf{x},5}[10]| = 12$. Contrarily to Example 40 and Fig. 5 where each prefix and suffix belong to two pairs, here one prefix (corresponding to the large interval) belongs to six pairs of the boundary and the other prefixes belong to one pair (or two for one short interval in the constellation on the right of Fig. 7).
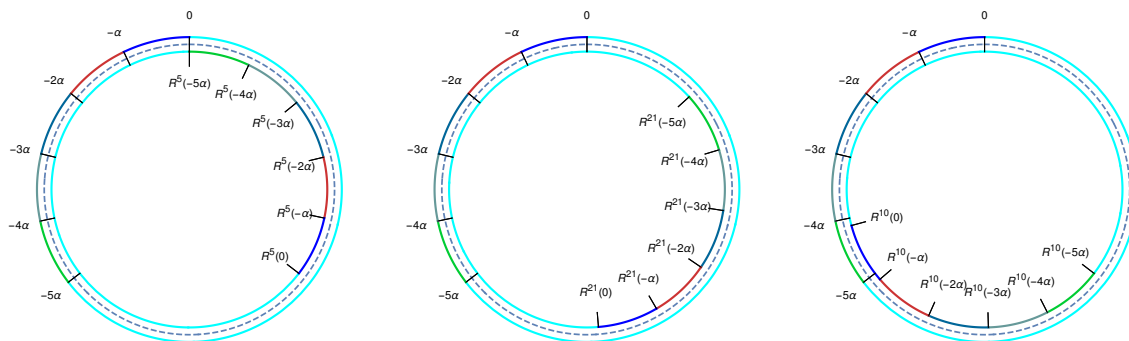

Figure 7: Constellations for $\alpha = (\pi - 3)/2$, $\ell = 5$, and $n = 5, 21, 10$.

Remark 42. 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 $\sigma\colon (0, 1) \mapsto (1, 0)$ and the identity mapping id; in this case $\partial_\sigma = \partial_{\mathrm{id}} = \{0, 1\} \times \{0, 1\}$. So two constellations inducing different maps in T_ℓ lead to the same set of boundary pairs. (See however Lemma 54.)

Definition 43. Let $\mathbf{r}$ be the rotation word defined by $\mathbf{r}[n] = \eta(R_\alpha^n(\alpha))$ for all $n \geqslant 0$, 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 of Theorem 37. Notice that by definition, the word $\mathbf{r}$ defined in Definition 39 is obtained by a sliding block code of length ℓ of the characteristic Sturmian word $\mathbf{s}_{\alpha,\alpha}$. We show that $T\partial_{\mathbf{s},\ell}$ is obtained from $\mathbf{r}$ by a sliding block code of length $\ell + 1$. The claim then follows since the composition of sliding block codes of length r and r', respectively, is a sliding block code of length $r + r' - 1$.

Let $n > \ell$. Consider the factor of length $\ell + 1$ of $\mathbf{r}$ occurring at position $m = n - \ell - 1$: by definition we have

$$\mathbf{r}[m]\mathbf{r}[m + 1] \cdots \mathbf{r}[m + \ell] = u_0 u_1 \cdots u_\ell$$

if and only if

$$u_{\ell - j} = \eta(R_\alpha^{m + \ell - j}(\alpha)) = \eta(R_\alpha^{m + \ell + 1}(\{-j\alpha\})),$$

for each $j \in \{0, \ldots, \ell\}$.

This is equivalent to $R^{m+\ell+1}(\{-j\alpha\}) \in I_{u_{\ell-j}}$ for each $j \in \{0, \ldots, \ell\}$. There thus exists a mapping $\sigma \in T_\ell$ such that $R_\alpha^{m+\ell+1}(\{-j\alpha\}) \in I_{w_{\sigma(j)}}$, whence $\partial_{\mathbf{s},\ell}[n] = \partial_{\mathbf{s},\ell}[m + \ell + 1] = \partial_\sigma$. We conclude that the factor of length $\ell + 1$ appearing at position m in $\mathbf{r}$ determines the boundary set $\partial_{\mathbf{s},\ell}[n]$. Letting the mapping $\mathcal{B} \colon \{0, \ldots, \ell\}^{\ell+1} \to T_\ell$ capture this relation, we may define an associated sliding block code $\mathfrak{B}$ of length $\ell + 1$ such that $\mathfrak{B}(\mathbf{r}) = T\partial_{\mathbf{s},\ell}$. $\qquad \square$

Example 44. We apply Theorem 37 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 $\mathbf{b} := \{0, 1\} \times \{0, 1\}$, and 00 to $\mathbf{a} := \{(0, 0), (0, 1), (1, 0)\}$. We may therefore define $\mathcal{B} \colon 01, 10 \mapsto \mathbf{b}, 00 \mapsto \mathbf{a}$ and the associated sliding block code $\mathfrak{B}$ of length 2; applying $\mathfrak{B}$ to $\mathbf{f}$, we get

$$\mathfrak{B}((01)(10)(00)(01)(10)(01)(10)(00)(01)(10)(00)(01)(10)(01)(10)(00)(01)(10)(01)\cdots)$$
$$= \mathbf{b}\ \ \mathbf{b}\ \ \mathbf{a}\ \ \mathbf{b}\ \ \mathbf{b}\ \ \mathbf{b}\ \ \mathbf{b}\ \ \mathbf{a}\ \ \mathbf{b}\ \ \mathbf{b}\ \ \mathbf{a}\ \ \mathbf{b}\ \ \mathbf{b}\ \ \mathbf{b}\ \ \mathbf{b}\ \ \mathbf{a}\ \ \mathbf{b}\ \ \mathbf{b}\ \ \mathbf{b}\cdots,$$

which indeed gives back Example 2 after prepending the letter $\mathbf{a}$.

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 45. *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.*

In what follows, for $x \in \mathbb{T}$, we define $\|x\| = \min\{x, 1 - x\}$, whence $\|x\| < 1/2$ for irrational x. It is not hard to show that 0^k or 1^k appears in a Sturmian word of slope α if and only if $k\|\alpha\| < 1$.

Proof of Proposition 45. Assume first that $\ell\|\alpha\| > 1/2$. Consider the set $R_\alpha^n(I_u) \cap I_v$ for some length-ℓ factors u, v, and $n \geqslant \ell$. We claim that it is an interval whenever it is non-empty. If it is not, then the intersection is a union of two intervals: without loss of generality $|I_u| > 1 - |I_v|$, and $R_\alpha^n(I_u)$ intersects I_v from both ends, but does not contain I_v

entirely. Notice that the intervals corresponding to length-ℓ factors have length at most $\|\alpha\|$ whenever $\ell\|\alpha\| > 1$. Since in that case we get the contradiction

$$|I_u| > 1 - |I_v| \geqslant 1 - \|\alpha\| > \|\alpha\|,$$

we must have $\ell\|\alpha\| < 1$. But now we know that the intervals have two admissible lengths, namely $\|\alpha\|$ and $1 - \ell\|\alpha\|$ (compare to the non-rotated points in Fig. 7 for an illustration). Now if $1 - \ell|\alpha\|$ is the largest of the two, we have a contradiction

$$|I_u| > 1 - |I_v| = 1 - \|\alpha\| \geqslant 1 - \ell\|\alpha\| = |I_u|.$$

Conversely, we get the contradiction

$$|I_u| > 1 - |I_v| = 1 - (1 - \ell\|\alpha\|) = \ell\|\alpha\| > 1/2 > \|\alpha\| = |I_u|.$$

We conclude that for any length-ℓ factors u, v of $\mathbf{s}$, the set $R_\alpha^n(I_u) \cap I_v$ is an interval or is empty. This implies that the boundary set $\partial_{\mathbf{x},\ell}[n]$ contains $2\ell + 2$ elements whenever $n > \ell$. Thus $\partial_{\mathbf{x},\ell}[\ell]$ occurs only once in the ℓ-boundary sequence due to a cardinality argument.

Assume then that $\ell\|\alpha\| < 1/2$. Without loss of generality we can assume $\alpha < 1/2$. It is straightforward to verify that

$$\partial_{\mathbf{x},\ell}[\ell] = \{(0^i 10^{\ell-i-1}, 0^\ell)\}_{i=0}^{\ell-1} \cup \{(0^\ell, 0^i 10^{\ell-i-1})\}_{i=0}^{\ell-1} \cup \{(0^\ell, 0^\ell)\}.$$

See, for instance, the first picture in Fig. 7. This same set is obtained for those n for which $0 < R_\alpha^n(\{-\ell\alpha\}) < R_\alpha^n(0) < 1 - \ell\|\alpha\|$: two of the $2\ell + 2$ intervals correspond to the boundary pair $(0^\ell, 0^\ell)$, namely the intervals $[0, R_\alpha^n(\{-\ell\alpha\}))$ and $[R_\alpha^n(0), 1 - \ell\|\alpha\|)$. Again see Fig. 7 for an illustration: $n = 21$ in the second picture satisfies the previous condition while $n = 10$ in the third picture does not. $\qquad\square$

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 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 37 and Proposition 45; 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 46. *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 $\ell \geqslant 2$. For any irrational number $\alpha \in (0, 1)$ there is a unique infinite continued fraction expansion

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

where $a_n \geqslant 1$ are integers for all $n \geqslant 1$. 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 $k \geqslant 1$ [2, Chap. 9]. The main result of this part is the following.

Proposition 47. *Let* $\mathbf{s}$ *be a Sturmian word of slope* $\alpha = [0; a_1 + 1, a_2, \ldots]$. *For each* $\ell \geqslant 2$, *there exists* $k_\ell \in \mathbb{N}$ *such that for any* $k \geqslant k_\ell$ *there is a morphism* $h_{k,\ell}$ *such that* $T\partial_{\mathbf{s},\ell} = h_{k,\ell}(\mathbf{s}_{\beta_k,\beta_k})$, *where* $\beta_k = [0; a_{k+1} + 1, a_{k+2}, \ldots]$.

Proof. Let $(S_j)_{j \geqslant -1}$ be the sequence associated to the slope α. Let then k be an integer such that $|S_k S_{k-1}| \geqslant 2\ell + 1$. Hence $\mathbf{s}_{\alpha,\alpha}$ is a product of S_k and S_{k-1}. It is now evident that with $\beta = [0; a_{k+1} + 1, a_{k+2}, \ldots]$, we have that $g(\mathbf{s}_{\beta,\beta}) = \mathbf{s}_{\alpha,\alpha}$, where g is defined by $g \colon 0 \mapsto S_k$, $1 \mapsto S_{k-1}$. Let $X = \mathrm{pref}_{2\ell-1}(S_k S_{k-1})$. We also have that $X = \mathrm{pref}_{2\ell-1}(S_{k-1}S_k)$, as $S_k S_{k-1}$ and $S_{k-1}S_k$ are known to differ in only the last two letters [2, Thm. 9.1.11]. We have that X is a prefix of both $S_k X$ and $S_{k-1}X$:

$$\mathrm{pref}_{2\ell-1}(S_k X) = \mathrm{pref}_{2\ell-1}(S_k\, \mathrm{pref}_{2\ell-1}(S_{k-1}S_k)) = \mathrm{pref}_{2\ell-1}(S_k S_{k-1}) = X$$

and

$$\mathrm{pref}_{2\ell-1}(S_{k-1}X) = \mathrm{pref}_{2\ell-1}(S_{k-1}\, \mathrm{pref}_{2\ell-1}(S_k S_{k-1})) = \mathrm{pref}_{2\ell-1}(S_{k-1}S_k) = X.$$

We define: $h \colon 0 \mapsto \mathfrak{B}(S_k X)$, $1 \mapsto \mathfrak{B}(S_{k-1}X)$, where $\mathfrak{B}$ is the sliding block code of length 2ℓ from Theorem 37 such that $\mathfrak{B}(\mathbf{s}_{\alpha,\alpha}) = T\partial_{\mathbf{s},\ell}$ (and T is the shift operator). Notice now that $\mathfrak{B}(uv) = \mathfrak{B}(u\, \mathrm{pref}_{2\ell-1}(v))\mathfrak{B}(v)$ for any sufficiently long word v (and u non-empty). Therefore

$$\begin{aligned}
h(\mathbf{s}_{\beta,\beta}) &= \mathfrak{B}(S_k X)^{a_{k+1}}\mathfrak{B}(S_{k-1}X)\mathfrak{B}(S_k X)^{a_{k+1}} \cdots \\
&= \mathfrak{B}(S_k^{a_{k+1}} S_{k-1} S_k^{a_{k+1}} \cdots) = \mathfrak{B}(\mathbf{s}_{\alpha,\alpha}) = T\partial_{\mathbf{s},\ell}. \qquad \square
\end{aligned}$$

We illustrate the above construction with a couple of examples for the benefit of the interested reader.

Example 48. Take the characteristic Sturmian word $\mathbf{s}$ of slope $\alpha = 1 - 1/\sqrt{3}$. The continued fraction expansion of α is $[0; 2, 2, 1, 2, 1, 2, 1, 2, 1, \ldots]$. Let $\ell = 2$. Then we have

$$S_{-1} = 1, \ S_0 = 0, \ S_1 = 01, \ S_2 = 01010, \ S_3 = 0101001, \ldots$$

Here $|S_2 S_1| = 7 \geqslant 5 = 2\ell + 1$. We have $X = \mathrm{pref}_3(S_2 S_1) = 010$. Then, defining $h \colon 0 \mapsto \mathfrak{B}(01010X) = 01234$, $1 \mapsto \mathfrak{B}(01X) = 01$, we find $T\partial_{\mathbf{s},\ell} = h(\mathbf{s})$ (see Lemma 54). Similarly, for $\ell = 3$ we have $|S_2 S_1| = 2\ell + 1$. Then $T\partial_{\mathbf{s},\ell} = h(\mathbf{s})$ when h is defined by $0 \mapsto \mathfrak{B}(01010X) = 01234$ and $1 \mapsto \mathfrak{B}(01X) = 56$, where $X = \mathrm{pref}_5(S_2 S_1) = S_2 = 01010$. Let then finally $\ell = 4$. Now we have $|S_2 S_1| = 7 < 9 = 2\ell + 1$, but $|S_3 S_2| = 12$. Hence $X = 0101001 = S_3$, and $T\partial_{\mathbf{s},\ell} = h(\mathbf{s}')$, where h is the morphism defined by

$$0 \mapsto \mathfrak{B}(0101001X) = 0123456, \ 1 \mapsto \mathfrak{B}(01010X) = 01278,$$

and $\mathbf{s}'$ is the characteristic Sturmian word whose slope β has continued fraction expansion $[0; 3, 1, 2, 1, 2, 1, 2, 1, 2, \ldots]$. One can verify that $\beta = 2 - \sqrt{3}$.

Example 49. The continued fraction expansion of $\alpha = \frac{1}{2}(\pi - 3)$ begins with

$$[0; 14, 7, 1, 586, 3, 1, 2, 1, 1, \ldots].$$

The construction in Proposition 47 gives $\partial_{\mathbf{s}_{\alpha,\alpha},2} = h(\mathbf{s}_{\beta,\beta})$, where $\beta = \frac{2}{\pi - 3} - 14$ has continued fraction expansion $[0; 7, 1, 586, 3, 1, 2, 1, 1, \ldots]$, and h is defined by $0 \mapsto 0^{10}1234$, $1 \mapsto 0$.

Example 50. Take the slope $\alpha = (3 - \sqrt{5})/2$; its continued fraction expansion is $[0; 2, 1, 1, 1, \ldots]$. Using the previous notation, $S_{-1} = 1$, $S_0 = 0$, and $S_{k+1} = S_k S_{k-1}$ for all $k \geqslant 0$. Then the sequence $(S_k)_{k \geqslant 0}$ converges to the Fibonacci word; the first few words in the sequence $(S_k)_{k \geqslant 0}$ are $0, 01, 010, 01001, 01001010$.

Now for any $\ell \geqslant 2$, 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 51. *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 [55] (see also [5]), characterizing those Sturmian words that are fixed by some non-trivial morphism, implies that if a Sturmian word of slope α is fixed by a non-trivial morphism, then so is the characteristic Sturmian word of slope α. Furthermore, the slope is characterized by the property that

$$\alpha = [0; 1, a_2, \overline{a_3, \ldots, a_r}]$$

with $a_r \geqslant a_2$ or

$$\alpha = [0; 1 + a_1, \overline{a_2, \ldots, a_r}]$$

with $a_r \geqslant a_1 \geqslant 1$ [15, 37] (see also [30, Thm. 2.3.25]). Here $\overline{x_1, \ldots, x_t}$ indicates the periodic tail of the infinite continued fraction expansion. As α is quadratic, it has an eventually periodic continued fraction expansion. There thus exist arbitrarily large k for which $\beta = [0; a_k + 1, a_{k+1}, \ldots]$ gives a characteristic Sturmian word of slope β which is the fixed point of a non-trivial morphism (it is of the latter form). Proposition 47 then posits that $T\partial_{\mathbf{s},\ell}$ is the morphic image of this word, and the claim follows (because prepending the letter $\partial_{\mathbf{s},\ell}[\ell]$ preserves morphicity [2, Thm. 7.6.3]). $\square$

Notice that given the morphism fixing a Sturmian word $\mathbf{s}$, one can compute (the continued fraction expansion of) the quadratic slope (and intercept) of $\mathbf{s}$ [54, 41, 29]. Furthermore, any (not necessarily pure) morphic Sturmian word has quadratic slope [1, 6], so in particular the boundary sequence of such a word is morphic.

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 [25], 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 [4]. 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 2.3.

We remark that it is unclear to us whether some of the results proved in this section could be proved automatically using the very recent tool `Pecan` developed in [36, 26].

4.3 Factor complexities of the extended boundary sequences

Definition 52. A word over an alphabet A is of *minimal complexity* if its factor complexity is $n + |A| - 1$ for all $n \geqslant 1$.

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 [38, 14, 20, 10, 17] for characterizations and generalizations.

The following proposition is almost immediate after the key Lemma 54.

Proposition 53. *Let* $\ell \geqslant 2$. *The ℓ-boundary sequence of a Sturmian word is a minimal complexity word (of complexity* $n \mapsto n + 2\ell$, $n \geqslant 1$).

Proof. Recall that $\partial_{\mathbf{s},\ell}$ is obtained by a coding of the 2ℓ-block coding of $\mathbf{s}_{\alpha,\alpha}$. The following lemma says that the coding is actually a bijection; in other words, a length-2ℓ factor of $\mathbf{s}$ uniquely determines a boundary set, or a letter, in the boundary sequence. We conclude that the factors of length $n + 2\ell - 1$ of $\mathbf{s}$ uniquely determine a factor of length n in the ℓ-boundary sequence. Since there are $n + 2\ell$ such factors of $\mathbf{s}$, the claim follows as the number of factors of length 2ℓ of $\mathbf{s}$, that is, the number of letters in $\partial_{\mathbf{s},\ell}$, is $2\ell + 1$. $\qquad\square$

Lemma 54. *Let* σ *and* $\sigma' \in T_\ell$, $\ell \geqslant 2$, *be distinct mappings both satisfying* (10) *(for different* n). *Then* $\partial_\sigma \neq \partial_{\sigma'}$.

Proof. Let σ (resp., σ') satisfy (10) with n (resp., m in place of n, $m \neq n$). Since $\sigma \neq \sigma'$, there exist $j \in \{0, \ldots, \ell\}$, and distinct factors $v, v' \in \mathrm{Fac}_\ell(\mathbf{s})$ such that $R_\alpha^n(\{-j\alpha\}) \in I_v$ and $R_\alpha^m(\{-j\alpha\}) \in I_{v'}$. To fix a rotation direction, assume without loss of generality that v' is lexicographically less than v, so $I_{v'}$ appears before I_v in clockwise order, starting from 0, in the 1-dimensional torus $\mathbb{T}$. The situation is depicted in Fig. 8: the interval I_v (resp., $I_{v'}$) is colored in orange (resp., dark red). Say that $\{-j\alpha\}$ is the starting point (in clockwise direction) of the interval I_u, and is the ending point of the interval I_w (again in clockwise direction); in particular, I_u and I_w are adjacent intervals. In particular, in Fig. 8, the interval $R_\alpha^n(I_u)$ in light turquoise (resp., $R_\alpha^m(I_u)$ in pink) appears after the interval $R_\alpha^n(I_w)$ in dark turquoise (resp., $R_\alpha^m(I_w)$ in purple) in clockwise order. We now have that ∂_σ contains (u, v) and (w, v), while $\partial_{\sigma'}$ contains (u, v') and (w, v'). Assume
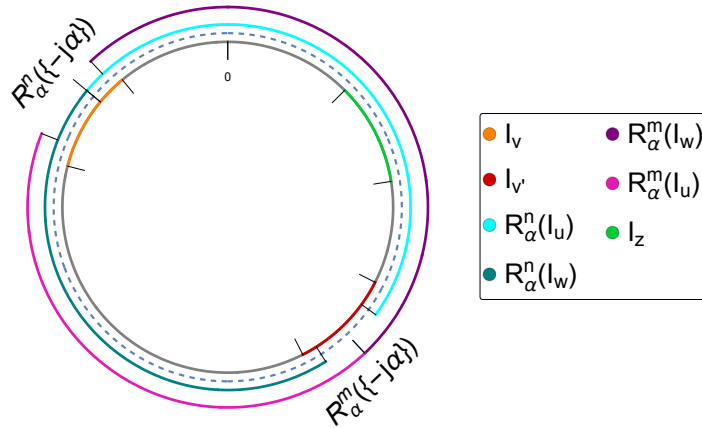
Figure 8: The situation depicted in the proof of Lemma 54.

towards a contradiction, that $\partial_\sigma = \partial_{\sigma'}$. Then we must have (u, v') and $(w, v') \in \partial_\sigma$ as well as (u, v) and $(w, v) \in \partial_{\sigma'}$. Now we have the following:

$$R_\alpha^n(I_u) \cap I_v \neq \emptyset \neq R_\alpha^n(I_u) \cap I_{v'}$$

(this is shown in Fig. 8 where the light turquoise interval intersects both the orange and dark red interval) and similarly

$$R_\alpha^n(I_w) \cap I_v \neq \emptyset \neq R_\alpha^n(I_w) \cap I_{v'}$$

(this is shown in Fig. 8 where the dark turquoise interval intersects both the orange and dark red interval). Since I_u and I_w are intervals, we see that $R^n(I_u)$ covers all intervals I_z between I_v and I'_v in clockwise order starting from the point $R_\alpha^n(\{-j\alpha\})$. Again, this is illustrated in Fig. 8 where an interval I_z is depicted in green. Similarly $R_\alpha^n(I_w)$ contains all intervals I_z between I_v and I'_v in anticlockwise order starting from the point $R_\alpha^n(\{-j\alpha\})$. The total number of the intermediate intervals I_z is $\ell + 1 - 2 = \ell - 1 \geqslant 1$, so assume without loss of generality that $R_\alpha^n(I_u)$ covers the interval I_z. In particular, this means that $(u, z) \in \partial_\sigma$. But, we have a symmetric situation as follows: the interval $R_\alpha^m(I_u)$ covers all intervals between $I_{v'}$ and I_v in clockwise order starting from $R_\alpha^m(\{-j\alpha\})$: these are the same intervals covered by $R^n(I_w)$. Since $(w, z) \notin \partial_\sigma$, we get the contradiction that $(u, z) \notin \partial_{\sigma'}$. This suffices for the claim. □

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

Proposition 55. *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*

$$n \mapsto \begin{cases} n + 1, & \text{if } n < 2r; \\ n + 2, & \text{otherwise.} \end{cases}$$

Proof. Without loss of generality, we assume that 00 appears in $\mathbf{s}$ and 11 does not. Let $\mathfrak{B}$ be the length-2 sliding block code from Theorem 37; it is not hard to show that $\mathfrak{B}$ is defined by $(00) \mapsto 0, (01), (10) \mapsto 1$. To prove the claim, we show that $\mathfrak{B}(\mathfrak{u}) = \mathfrak{B}(\mathfrak{v})$ with $\mathfrak{u} \neq \mathfrak{v}$ if and only if $\mathfrak{u}$ is a prefix of $(01)^{\mathfrak{r}}$ and $\mathfrak{v}$ is a prefix of $(10)^{\mathfrak{r}}$ (assuming $|\mathfrak{u}|, |\mathfrak{v}| \geqslant 2$). This is enough since, as in the proof of Proposition 53, a factor of length $\mathfrak{n} + 1$ of $\mathbf{s}$ corresponds to a factor of length $\mathfrak{n}$ of $\partial_{\mathbf{s}}$.

Observe that if $\mathfrak{u}$ is a prefix of $(01)^{\mathfrak{r}}$ and $\mathfrak{v}$ is a prefix of $(10)^{\mathfrak{r}}$, then $\mathfrak{B}(\mathfrak{u}) = \mathfrak{B}(\mathfrak{v}) = 1^{|\mathfrak{u}|-1}$. Let us show the converse by induction on the length of $\mathfrak{u}, \mathfrak{v}$, and hence assume that $\mathfrak{B}(\mathfrak{u}) = \mathfrak{B}(\mathfrak{v})$ with $\mathfrak{u} \neq \mathfrak{v}$. If $|\mathfrak{u}| = 2 = |\mathfrak{v}|$, the claim is clear. Assume then that $|\mathfrak{u}|, |\mathfrak{v}| > 2$. If $\mathfrak{u}$ and $\mathfrak{v}$ begin with the same letter, then their second letter must be equal, because otherwise $\mathfrak{B}(\mathfrak{u})$ begins with 0 and $\mathfrak{B}(\mathfrak{v})$ with 1 or vice versa. So write $\mathfrak{u} = \mathfrak{abu}'$ and $\mathfrak{v} = \mathfrak{abv}'$ for some letters $\mathfrak{a}, \mathfrak{b} \in \{0, 1\}$ and some binary words $\mathfrak{u}', \mathfrak{v}'$. Since the words $\mathfrak{bu}'$ and $\mathfrak{bv}'$ are shorter and distinct, and have equal $\mathfrak{B}$-images, the induction hypothesis implies that one is a prefix of $(01)^{\mathfrak{r}}$ and the other a prefix of $(10)^{\mathfrak{r}}$. This is, of course, impossible. We conclude that the words $\mathfrak{u}$ and $\mathfrak{v}$ begin with distinct letters. Without loss of generality, suppose that $\mathfrak{u}$ begins with 0 and $\mathfrak{v}$ with 1. Since 11 does not appear in $\mathbf{s}$, we deduce that $\mathfrak{v}$ begins with 10, hence $\mathfrak{B}(\mathfrak{v})$ begins with 1. Therefore $\mathfrak{u}$ must begin with 01 for $\mathfrak{B}(\mathfrak{u})$ to begin with 1. Removing the first letter of $\mathfrak{u}$ and $\mathfrak{v}$ allows us to use induction to complete the claim. $\square$

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

5   Conclusions

There is no particular reason to consider boundary pairs of equal length. One may just as well define the $(\mathfrak{k}, \ell)$-boundary sequence in an analogous manner. All the results appearing in Sections 2.3 and 3 can be extended straightforwardly to account for this seemingly more general notion. The methods used in Section 4 can also be adapted to deal with $(\mathfrak{k}, \ell)$-boundary sequences straightforwardly.

Acknowledgments

References

[1] Jean-Paul Allouche, Julien Cassaigne, Jeffrey Shallit, and Luca Q. Zamboni. A taxonomy of morphic sequences, 2017. `arXiv:1711.10807`.

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

[3] Tom M. Apostol. *Introduction to analytic number theory*. Undergraduate Texts in Mathematics. Springer-Verlag, New York-Heidelberg, 1976.

[4] 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`.

[5] 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`.

[6] Valérie Berthé, Charles Holton, and Luca Q. Zamboni. Initial powers of Sturmian sequences. *Acta Arithmetica*, 122(4):315–347, 2006. `doi:10.4064/aa122-4-1`.

[7] 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`.

[8] 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`.

[9] Véronique Bruyère, Georges Hansel, Christian Michaux, and Roger Villemaire. Logic and $\mathfrak{p}$-recognizable sets of integers. *Bulletin of the Belgian Mathematical Society. Simon Stevin*, 1(2):191–238, 1994. Journées Montoises (Mons, 1992). `doi:10.36045/bbms/1103408547`.

[10] 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.

[11] 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`.

[12] 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`.

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

[14] Ethan M. Coven. Sequences with minimal block growth II. *Mathematical Systems Theory*, 8:376–382, 1974. `doi:10.1007/BF01780584`.

[15] 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`.

[16] 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`.

[17] 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`.

[18] Fabien Durand and Michel Rigo. Syndeticity and independent substitutions. *Advances in Applied Mathematics*, 42(1):1–22, 2009. `doi:10.1016/j.aam.2008.02.001`.

[19] 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`.

[20] 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`.

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

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

[23] 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`.

[24] 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`.

[25] 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`.

[26] 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`.

[27] 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`.

[28] Pierre B. A. Lecomte and Michel Rigo. Numeration systems on a regular language. *Theory of Computing Systems*, 34(1):27–44, 2001. `doi:10.1007/s002240010014`.

[29] Jana Lepšová, Edita Pelantová, and Štěpán Starosta. On a faithful representation of Sturmian morphisms. *European Journal of Combinatorics*, 110:Paper No. 103707, 20, 2023. `doi:10.1016/j.ejc.2023.103707`.

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

[31] Xiao-Tao Lü, Jin Chen, Zhi-Xiong Wen, and Wen Wu. On the 2-binomial complexity of the generalized Thue-Morse words, 2021. Preprint. `arXiv:2112.05347`.

[32] 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.

[33] 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`.

[34] Hamoon Mousavi. Walnut prover, 2016. `https://github.com/hamousavi/Walnut`, `https://cs.uwaterloo.ca/~shallit/walnut.html`.

[35] 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`.

[36] Reed Oei, Dun Ma, Christian Schulz, and Philipp Hieronymi. Pecan: An automated theorem prover for automatic sequences using Büchi automata, 2021. `arXiv:2102.01727`.

[37] 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`.

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

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

[40] 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`.

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

[42] 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`.

[43] Michel Rigo. Numeration systems on a regular language: Arithmetic operations, recognizability and formal power series. *Theoretical Computer Science*, 269(1-2):469–498, 2001. `doi:10.1016/S0304-3975(01)00184-0`.

[44] Michel Rigo. Construction of regular languages and recognizability of polynomials. *Discrete Mathematics*, 254(1-3):485–496, 2002. `doi:10.1016/S0012-365X(01)00377-6`.

[45] 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.

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

[47] 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`.

[48] Michel Rigo, Manon Stipulanti, and Markus A. Whiteland. On Extended Boundary Sequences of Morphic and Sturmian Words. In Stefan Szeider, Robert Ganian, and Alexandra Silva, editors, *47th International Symposium on Mathematical Foundations of Computer Science (MFCS 2022)*, volume 241 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:16, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. `doi:10.4230/LIPIcs.MFCS.2022.79`.

[49] Michel Rigo, Manon Stipulanti, and Markus A. Whiteland. Characterizations of families of morphisms and words via binomial complexities. *European Journal of Combinatorics*, 2023. To appear. `arXiv:2201.04603`.

[50] Klaus Schmidt. On periodic expansions of Pisot numbers and Salem numbers. *Bulletin of the London Mathematical Society*, 12:269–278, 1980. `doi:10.1112/blms/12.4.269`.

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

[52] Jeffrey Shallit. *The Logical Approach to Automatic Sequences: Exploring Combinatorics on Words with Walnut.* London Mathematical Society Lecture Note Series. Cambridge University Press, 2022. `doi:10.1017/9781108775267`.

[53] Neil Sloane. The On-Line Encyclopedia of Integer Sequences. http://oeis.org.

[54] 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`.

[55] 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.