On Infinite Prefix Normal Words

Ferdinando Cicalese, Zsuzsanna Lipták, and Massimiliano Rossi

Dipartimento di Informatica, University of Verona
Strada le Grazie, 15, 37134 Verona, Italy
{ferdinando.cicalese,zsuzsanna.liptak,massimiliano.rossi_01}@univr.it

Abstract. Prefix normal words are binary words that have no factor with more 1s than the prefix of the same length. Finite prefix normal words were introduced in [Fici and Lipták, DLT 2011]. In this paper, we study infinite prefix normal words and explore their relationship to some known classes of infinite binary words. In particular, we establish a connection between prefix normal words and Sturmian words, between prefix normal words and abelian complexity, and between prefix normality and lexicographic order.

keywords combinatorics on words, prefix normal words, infinite words, Sturmian words, abelian complexity, paperfolding word, Thue-Morse sequence, lexicographic order

1 Introduction

Prefix normal words are binary words where no factor has more 1s than the prefix of the same length. As an example, the word 11100110101 is prefix normal, while 11100110110 is not, since it has a factor of length 5 with 4 1s, while the prefix of length 5 has only 3 1s. Finite prefix normal words were introduced in [12] and further studied in [5, 6, 9, 22].

The original motivation for studying prefix normal words comes from the problem of *Indexed Binary Jumbled Pattern Matching* [1, 4, 8, 10, 13]. Given a finite word s of length n, construct an index in such a way that the following type of queries can be answered efficiently: For two integers $x, y \geq 0$, does s have a factor with x 1s and y 0s? As shown in [12], prefix normal words can be used for constructing such an index, via so-called *prefix normal forms*.

Prefix normal words have been shown to form bubble languages [5, 20, 21], a family of binary languages with efficiently generable combinatorial Gray codes; they have connections to the Binary Reflected Gray Code [22]; and they have recently found application to a certain class of graphs [3]. Indeed, three sequences related to prefix normal words are present in the On-Line Encyclopedia of Integer Sequences (OEIS [23]): A194850 (the number of prefix normal words of length n), A238109 (a list of prefix normal words over the alphabet $\{1, 2\}$), and A238110 (equivalence class sizes of words with the same prefix normal form).

In [9], we introduced infinite prefix normal words and analyzed a particular procedure that, given a finite prefix normal word, extends it while preserving

the prefix normality property. We showed that the resulting infinite word is ultimately periodic. In this paper, we present a more comprehensive study of infinite prefix normal words, covering several classes of known and well studied infinite words. We now give a quick tour of the paper.

There exist periodic, ultimately periodic, and aperiodic infinite prefix normal words (for precise definitions, see Section 2): for example, the periodic words $0^\omega, 1^\omega$, and $(10)^\omega$ are prefix normal; the ultimately periodic word $1(10)^\omega$ is prefix normal; and so is the aperiodic word $10100100010000\cdots = \lim_{n\to\infty} 1010^2\cdots 10^n$. In Section 3, we fully characterize periodic and ultimately periodic words in terms of their minimum density, a parameter introduced in [9].

Regarding aperiodic words, we show that a Sturmian word w is prefix normal if and only if $w = 1c_\alpha$ for some α, where c_α is the characteristic word of slope α (Theorem 2). The Fibonacci word $f = 0100101001001010010100100101001001\cdots$ is thus not prefix normal, but we can turn it into a prefix normal word by prepending a 1, i.e. the word $1f$ is prefix normal. We show in fact that every Sturmian word w can be turned into a prefix normal word by prepending a fixed number of 1s, which only depends on the slope of w. This follows from a more general result regarding c-balanced words (Lemma 7).

The Thue-Morse word $tm = 01101001100101101001011001101001\cdots$ is not prefix normal, but $11tm$ is. However, the binary Champernowne word, which is constructed by concatenating the binary expansions of the integers in ascending order, namely $c = 0110111001011101111000100110101011\cdots$ is not prefix normal and cannot be turned into a prefix normal word by prepending a finite number of 1s, because c has arbitrarily long runs of 1s.

One might be tempted to conclude that every word with bounded abelian complexity can be turned into a prefix normal word by prepending a fixed number of 1s, as is the case for the words above: f has abelian complexity constant 2, tm has abelian complexity bounded by 3, and c has unbounded abelian complexity. This is not the case, as we will see in Section 5.

We further show in Section 5 that the notion of prefix normal *forms* from [12] can be extended to infinite words. As in the finite case, these can be used to encode the abelian complexity of the original word. The study of abelian complexity of infinite words was initiated in [18], and continued e.g. in [2,7,14,16,25]. We establish a close relationship between the abelian complexity and the prefix normal forms of w. We demonstrate how this close connection can be used to derive results about the prefix normal forms of a word w. In some cases, such as for Sturmian words and words which are morphic images under the Thue-Morse morphism, we are able to explicitly give the prefix normal forms of the word. Conversely, knowing its prefix normal forms allows us to compute the abelian complexity function of a word.

Another class of well-known binary words are Lyndon words. Notice that the prefix normal condition is different from the Lyndon condition[1]: For finite words, there are words which are both Lyndon and prefix normal (e.g. 110010),

___
[1] For ease of presentation, we use Lyndon to mean lexicographically *greatest* among its conjugates; this is equivalent to the usual definition up to renaming characters.

words which are Lyndon but not prefix normal (11100110110), words which are prefix normal but not Lyndon (110101), and words which are neither (101100). In the final part of the paper, we will put infinite prefix normal words and their prefix normal forms in the context of lexicographic orderings, and compare them to infinite Lyndon words [24] and the max- and min-words of [17].

The paper is organized as follows. In Section 2, we introduce our terminology and give some simple facts about prefix normal words. In Section 3, we introduce the notion of *minimum density* and show its utility in dealing with certain prefix normal words. In Section 4, we study the relationship of Sturmian and prefix normal words. Section 5 treats prefix normal forms and their close connection to abelian complexity, and in Section 6 we study the relationship with lexicographic order. All proofs were moved to the Appendix.

2   Basics

In our definitions and notations, we follow mostly [15], wherever possible. A finite (resp. infinite) binary word w is a finite (resp. infinite) sequence of elements from $\{0, 1\}$. Thus an infinite word is a mapping $w : \mathbb{N} \to \{0, 1\}$, where $\mathbb{N}$ denotes the set of positive integers. We denote the i'th character of w by w_i. Note that we index words starting from 1. If w is finite, then its length is denoted by $|w|$. The empty word, denoted ε, is the unique word of length 0. The set of binary words of length n is denoted by $\{0, 1\}^n$, the set of all finite words by $\{0, 1\}^* = \cup_{n \geq 0} \{0, 1\}^n$, and the set of infinite binary words by $\{0, 1\}^\omega$. For a finite word $u = u_1 \cdots u_n$, we write $u^{\text{rev}} = u_n \cdots u_1$ for the reverse of u, and $\overline{u} = \overline{u}_1 \cdots \overline{u}_n$ for the complement of u, where $\overline{a} = 1 - a$.

For two words u, v, where u is finite and v is finite or infinite, we write uv for their concatenation. If $w = uxv$, then u is called a prefix, x a factor (or substring), and v a suffix of w. We denote the set of factors of w by $Fct(w)$ and its prefix of length i by $\text{pref}_w(i)$, where $\text{pref}_w(0) = \varepsilon$. For a finite word u, we write $|u|_1$ for the number of 1s, and $|u|_0$ for the number of 0s in u, and refer to $|u|_1$ as the *weight* of u. The *Parikh vector* of u is $pv(u) = (|u|_0, |u|_1)$. A word w is called *balanced* if for all $u, v \in Fct(w)$, $|u| = |v|$ implies $||u|_1 - |v|_1| \leq 1$, and *c-balanced* if $|u| = |v|$ implies $||u|_1 - |v|_1| \leq c$.

For an integer $k \geq 1$ and $u \in \{0, 1\}^n$, u^k denotes the kn-length word $uuu \cdots u$ (k-fold concatenation of u) and u^ω the infinite word $uuu \cdots$. An infinite word w is called *periodic* if $w = u^\omega$ for some non-empty word u, and *ultimately periodic* if it can be written as $w = vu^\omega$ for some v and non-empty u. A word that is neither periodic nor ultimately periodic is called *aperiodic*. We set $0 < 1$ and denote by $\leq_{\text{lex}}$ the *lexicographic order* between words, i.e. $u \leq_{\text{lex}} v$ if u is a prefix of v or there is an index $i \geq 1$ s.t. $u_i < v_i$ and $\text{pref}_u(i - 1) = \text{pref}_v(i - 1)$.

For an operation op : $\{0, 1\}^* \to \{0, 1\}^*$, we denote by $\text{op}^{(i)}$ the ith iteration of op; $\text{op}^*(w) = \{\text{op}^{(i)}(w) \mid i \geq 1\}$; and $\text{op}^\omega(w) = \lim_{i \to \infty} \text{op}^{(i)}(w)$, if it exists.

Definition 1 (Prefix weight, prefix density, maximum and minimum 1s and 0s functions). *Let w be a (finite or infinite) binary word. We define the following functions:*

- $P_w(i) = |\operatorname{pref}_w(i)|_1$, *the* weight *of the prefix of length i,*
- $D_w(i) = P_w(i)/i$, *the* density *of the prefix of length i,*
- $F_w^1(i) = \max\{|u|_1 : u \in Fct(w), |u| = i\}$ *and* $f_w^1(i) = \min\{|u|_1 : u \in Fct(w), |u| = i\}$, *the maximum resp. minimum number of 1s in a factor of length i,*
- $F_w^0(i) = \max\{|u|_0 : u \in Fct(w), |u| = i\}$ *and* $f_w^0(i) = \min\{|u|_0 : u \in Fct(w), |u| = i\}$, *the maximum resp. minimum number of 0s in a factor of length i.*

Note that in the context of succinct indexing, the function $P_w(i)$ is often called $rank_1(w, i)$. We are now ready to define prefix normal words.

Definition 2 (Prefix normal words). *A (finite or infinite) binary word w is called 1-prefix normal, if $P_w(i) = F_w^1(i)$ for all $i \geq 1$ (for all $1 \leq i \leq |w|$ if w is finite). It is called 0-prefix normal if $i - P_w(i) = F_w^0(i)$ for all $i \geq 1$ (for all $1 \leq i \leq |w|$ if w is finite). We denote the set of all finite 1-prefix normal words by $\mathcal{L}_{\mathrm{fin}}$, the set of all infinite 1-prefix normal words by $\mathcal{L}_{\mathrm{inf}}$, and $\mathcal{L} = \mathcal{L}_{\mathrm{fin}} \cup \mathcal{L}_{\mathrm{inf}}$.*

In other words, a word is prefix normal (i.e. 1-prefix normal) if no factor has more 1s than the prefix of the same length. Note that unless further specified, by prefix normal we mean 1-prefix normal. Given a binary word w, we say that a factor u of w *satisfies the prefix normal condition* if $|u|_1 \leq P_w(|u|)$.

Example 1. The word 110100110110 is not prefix normal since the factor 11011 has four 1s, which is more than in the prefix 11010 of length 5. The word 110100110010, on the other hand, is prefix normal. The infinite word $(11001)^\omega$ is not prefix normal, because of the factor 111, but the word $(11010)^\omega$ is.

The following facts about infinite prefix normal words are immediate.

Lemma 1. *1. For all $u \in \mathcal{L}_{\mathrm{fin}}$, the word $w = u0^\omega \in \mathcal{L}_{\mathrm{inf}}$.*
2. Let $w \in \{0, 1\}^\omega$. Then $w \in \mathcal{L}$ if and only if for all $i \geq 1$, $\operatorname{pref}_w(i) \in \mathcal{L}$.

Definition 3 (Minimum density, minimum-density prefix, slope).
Let $w \in \{0, 1\}^* \cup \{0, 1\}^\omega$. Define the minimum density of w as $\delta(w) = \inf\{D_w(i) \mid 1 \leq i\}$. If this infimum is attained somewhere, then we also define $\iota(w) = \min\{j \geq 1 \mid \forall i : D_w(j) \leq D_w(i)\}$ and $\kappa(w) = P_w(\iota(w))$. We refer to $\operatorname{pref}_w(\iota(w))$ as the minimum-density prefix, the shortest prefix with density $\delta(w)$. For an infinite word w, we define the slope of w as $\lim_{i \to \infty} D_w(i)$, if this limit exists.

Remark 1. Note that $\iota(w)$ is always defined for finite words, while for infinite words, a prefix which attains the infimum may or may not exist. We note further that density and slope are different properties of (infinite) binary words. In particular, while $\delta(w)$ exists for every w, the limit $\lim_{i \to \infty} D_w(i)$ may not exist,

i.e., w may or may not have a slope. As an example, consider $w = v_0 v_1 v_2 \cdots$, where for each i, $v_i = 1^{2^i} 0^{2^i}$; then $\delta(w) = 1/2$ and $\lim_{i \to \infty} D_w(i)$ does not exist, since $D_w(i)$ has an infinite subsequence constant $1/2$, and another which tends to $2/3$. But even for words w whose slope is defined, it can be different from $\delta(w)$. If w has slope α, then $\alpha = \delta(w)$ if and only if for all i, $D_w(i) \geq \alpha$. For instance, the infinite word 01^ω has slope 1 but its minimum density is 0. On the other hand, the infinite word $1(10)^\omega$ has both slope and minimum density $1/2$.

3   A characterization of periodic and aperiodic prefix normal words with respect to minimum density

In [9], we introduced an operation which takes a finite prefix normal word w containing at least one 1 and *extends* it by a run of 0s followed by a new 1, in such a way that this 1 is placed in the first possible position without violating prefix normality. This operation, called flipext, leaves the minimum density invariant.

Definition 4 ([9] Operation flipext**).** *Let $w \in \mathcal{L}_{\mathrm{fin}} \setminus \{0\}^*$. Define* flipext$(w)$ *as the finite word $w0^k1$, where $k = \min\{j \mid w0^j1 \in \mathcal{L}\}$. We further define the infinite word $v = $ flipext$^\omega(w) = \lim_{i \to \infty}$ flipext$^{(i)}(w)$.*

Proposition 1 ([9]). *Let $w \in \mathcal{L}_{\mathrm{fin}} \setminus \{0\}^*$ and $v \in $ flipext$^*(w) \cup \{$flipext$^\omega(w)\}$. Then it holds that $\delta(v) = \delta(w)$, and as a consequence, $\iota(v) = \iota(w)$ and $\kappa(v) = \kappa(w)$. Moreover, $D_v(k \cdot \iota(w)) = \delta(w)$ for each $k \geq 1$.*

The following result shows that every ultimately periodic infinite prefix normal word has rational minimum density.

Lemma 2. *Let v be an infinite ultimately periodic binary word with minimum density $\delta(v) = \alpha$. Then $\alpha \in \mathbb{Q}$.*

Next we show that conversely, for every $\alpha \in (0, 1)$, both rational and irrational, there is an aperiodic prefix normal word with minimum density α.

Lemma 3. *Fix $\alpha \in (0, 1)$, and let $(a_n)_{n \in \mathbb{N}}$ be a strictly decreasing infinite sequence of rational numbers from $(0, 1)$ converging to α. For each $i = 1, 2, \ldots$, let the binary word $v^{(i)}$ be defined by*

$$v^{(i)} = \begin{cases} 1^{\lceil 10a_1 \rceil} 0^{10 - \lceil 10a_1 \rceil} & i = 1 \\ \mathrm{pref}_{\mathrm{flipext}^\omega(v^{(i-1)})}(k_i|v^{(i-1)}|)0^{\ell_i} & i > 1 \end{cases} \qquad \ell_i = \begin{cases} 10 - \lceil 10a_1 \rceil & i = 1 \\ \left\lfloor k_i \left(\frac{|v^{(i-1)}|_1 - a_i|v^{(i-1)}|}{a_i} \right) \right\rfloor & i > 1, \end{cases}$$

and k_i is the smallest integer greater than one such that $\ell_i > \ell_{i-1}$. Then $v = \lim_{i \to \infty} v^{(i)}$ is an aperiodic infinite prefix normal word such that $\delta(v) = \alpha$.

Summarizing, we have shown the following result.

Theorem 1. *For every $\alpha \in (0, 1)$ (rational or irrational) there is an infinite aperiodic prefix normal word of minimum density α. On the other hand, for every ultimately periodic infinite prefix normal word w the minimum density $\delta(w)$ is a rational number.*

4   Sturmian words and prefix normal words

The results of the previous section show that there is a relationship between the rationality or irrationality of the minimum density of an infinite prefix normal word and its aperiodic or periodic behaviour. This is reminiscent of the characterization of Sturmian words in terms of the slope. Led by this analogy, in this section we provide a complete characterization of Sturmian words which are prefix normal. We refer the interested reader to [15], chapter 2, for a comprehensive treatment of Sturmian words. Here we briefly recall some facts we will need, starting with two equivalent definitions of Sturmian words.

Definition 5 (Sturmian words). *Let $w \in \{0,1\}^\omega$. Then w is called* Sturmian *if it is balanced and aperiodic.*

Definition 6 (Mechanical words). *Given two real numbers $0 \leq \alpha \leq 1$ and $0 \leq \tau < 1$, the* lower mechanical word *$s_{\alpha,\tau} = s_{\alpha,\tau}(1)\,s_{\alpha,\tau}(2)\,\cdots$ and the* upper mechanical word *$s'_{\alpha,\tau} = s'_{\alpha,\tau}(1)\,s'_{\alpha,\tau}(2)\,\cdots$ are given by*

$$
\begin{aligned}
s_{\alpha,\tau}(n) &= \lfloor \alpha n + \tau \rfloor - \lfloor \alpha(n-1) + \tau \rfloor \\
s'_{\alpha,\tau}(n) &= \lceil \alpha n + \tau \rceil - \lceil \alpha(n-1) + \tau \rceil
\end{aligned}
\qquad (n \geq 1).
$$

Then α is called the slope *and τ the* intercept *of $s_{\alpha,\tau}, s'_{\alpha,\tau}$. A word w is called* mechanical *if $w = s_{\alpha,\tau}$ or $w = s'_{\alpha,\tau}$ for some α, τ. It is called* rational mechanical *(resp.* irrational mechanical*) if α is rational (resp. irrational).*

Fact 1 (Some facts about Sturmian words [15])   *1. An infinite binary word is Sturmian if and only if it is irrational mechanical.*
2. For $\tau = 0$, and α irrational, there exists a word c_α, called the characteristic *word with slope α, s.t. $s_{\alpha,0} = 0c_\alpha$ and $s'_{\alpha,0} = 1c_\alpha$. This word c_α is a Sturmian word itself, with both slope and intercept α.*
3. For two Sturmian words w, v with the same slope, we have $Fct(w) = Fct(v)$.

4.1 From flipext to lazy-α-flipext. Recall the operation flipext(w) defined above (Def. 4). We now define a different operation that, given a prefix normal word w, extends it by adding 0s as long as the minimum density of the resulting word is not smaller than $\delta(w)$, and only then adding a 1. We show that this operation preserves the prefix normality of the word. The operation lazy-α-flipext is then applied to show that, by extending a prefix normal word w of minimum density at least α, in the same way as we compute the upper mechanical word of slope α, we obtain an infinite prefix normal word with prefix w.

Definition 7. *Let $\alpha \in (0,1]$ and $w \in \mathcal{L}_{\text{fin}}$ with $\delta(w) \geq \alpha$. Define* lazy-α-flipext(w) *as the finite word $w0^k 1$ where $k = \max\{j \mid \delta(w0^j) \geq \alpha\}$. We further define the infinite word $v = $ lazy-α-flipext$^\omega(w) = \lim_{i \to \infty}$ lazy-α-flipext$^{(i)}(w)$.*

Example 2. Let $w = 111$ and let $\alpha = \sqrt{2}-1$, then lazy-α-flipext$(w) = 11100001$, since $\delta(1110000) = 3/7 \geq \alpha$ and $\delta(11100000) = 3/8 < \alpha$; and lazy-$\alpha$-flipext$^{(2)}(w) = 1110000101$, since $\delta(111000010) = 4/9 \geq \alpha$ and $\delta(1110000100) = 2/5 < \alpha$.

Lemma 4. *Let $\alpha \in (0,1]$. For every $w \in \mathcal{L}_{\text{fin}}$ with $\delta(w) \geq \alpha$, the word $v = \text{lazy-}\alpha\text{-flipext}(w)$ is also prefix normal, with $\delta(v) \geq \alpha$.*

Corollary 1. *Let $\alpha \in (0,1]$, $w \in \mathcal{L}_{\text{fin}}$ with $\delta(w) \geq \alpha$. Then $v = \text{lazy-}\alpha\text{-flipext}^{\omega}(w)$ is an infinite prefix normal word and $\delta(v) = \alpha$.*

We now show that the word $\text{lazy-}\alpha\text{-flipext}^{\omega}(1)$ coincides with the upper mechanical word $s'_{\alpha,0}$, which also implies that $s'_{\alpha,0}$ is prefix normal.

Lemma 5. *Fix $\alpha \in (0,1]$ and let $v = \text{lazy-}\alpha\text{-flipext}^{\omega}(1)$. Let $s = s'_{\alpha,0}$ be the upper mechanical word of slope α and intercept 0. Then $v = s$.*

Corollary 2. *For $\alpha \in (0,1]$, the word $s'_{\alpha,0}$ is prefix normal and $\delta(s'_{\alpha,0}) = \alpha$.*

The following theorem fully characterizes prefix normal Sturmian words.

Theorem 2. *A Sturmian word s of slope α is prefix normal if and only if $s = 1c_\alpha$, where c_α is the characteristic Sturmian word with slope α.*

5 Prefix normal words, prefix normal forms, and abelian complexity

Given an infinite word w, the *abelian complexity* function of w, denoted ψ_w, is given by $\psi_w(n) = |\{pv(u) \mid u \in Fct(w), |u| = n\}|$, the number of Parikh vectors of n-length factors of w. A word w is said to have bounded abelian complexity if there exists a c s.t. for all n, $\psi_w(n) \leq c$. Note that a binary word is c-balanced if and only if its abelian complexity is bounded by $c + 1$. We denote the set of Parikh vectors of factors of a word w by $\Pi(w) = \{pv(u) \mid u \in Fct(w)\}$. Thus, $\psi_w(n) = \Pi(w) \cap \{(x,y) \mid x + y = n\}$. In this section, we study the connection between prefix normal words and abelian complexity.

5.1 Balanced and c-balanced words. Based on the examples in the introduction, one could conclude that any word with bounded abelian complexity can be turned into a prefix normal word by prepending a fixed number of 1s. However, consider the word $w = 01^\omega$, which is balanced, i.e. its abelian complexity function is bounded by 2. It is easy to see that $1^k w \notin \mathcal{L}$ for every $k \in \mathbb{N}$.

Sturmian words are precisely the words which are aperiodic and whose abelian complexity is constant 2 [18]. For Sturmian words, it is always possible to prepend a finite number of 1s to get a prefix normal word, as we will see next. Recall that for a Sturmian word w, at least one of $0w$ and $1w$ is Sturmian, with both being Sturmian if and only if w is characteristic [15].

Lemma 6. *Let w be a Sturmian word. Then*
    1. $1w \in \mathcal{L}$ if and only if $0w$ is Sturmian,
    2. if $0w$ is not Sturmian, then $1^n w \in \mathcal{L}$ for $n = \lceil 1/(1 - \alpha) \rceil$.

Lemma 7. *Let w be a c-balanced word. If there exists a positive integer n s.t. $1^n \notin Fct(w)$, then the word $z = 1^{nc} w$ is prefix normal.*

In particular, Lemma 7 implies that any c-balanced word with infinitely many 0s can be turned into a prefix normal word by prepending a finite number of 1s, since such a word cannot have arbitrarily long runs of 1s. Note, however, that the number of 1s to prepend from Lemma 7 is not tight, as can be seen e.g. from the Thue-Morse word tm: the longest run of 1s in tm is 2 and tm is 2-balanced, but $11\mathsf{tm}$ is prefix normal, as will be shown in the next section (Lemma 10).

5.2 Prefix normal forms and abelian complexity. Recall that for a word w, $F_w^a(i)$ is the maximum number of a's in a factor of w of length i, for $a \in \{0, 1\}$.

Definition 8 (Prefix normal forms). *Let $w \in \{0, 1\}^\omega$. Define the words w' and w'' by setting, for $n \geq 1$, $w'_n = F_w^1(n) - F_w^1(n-1)$ and $w''_n = \overline{F_w^0(n) - F_w^0(n-1)}$. We refer to w' as the* prefix normal form of w w.r.t. 1 *and to w'' as the* prefix normal form of w w.r.t. 0, *denoted* $\mathrm{PNF}_1(w)$ *resp.* $\mathrm{PNF}_0(w)$.

In other words, $\mathrm{PNF}_1(w)$ is the sequence of first differences of the maximum-1s function F_w^1 of w. Similarly, $\mathrm{PNF}_0(w)$ can be obtained by complementing the sequence of first differences of the maximum-0s function F_w^0 of w. Note that for all n and $a \in \{0, 1\}$, either $F_w^a(n + 1) = F_w^a(n)$ or $F_w^a(n + 1) = F_w^a(n) + 1$, and therefore w' and w'' are words over the alphabet $\{0, 1\}$. In particular, by construction, the two prefix normal words allow us to recover the maximum-1s and minimum-1s functions of w:

Observation 1 *Let w be an infinite binary word and $w' = \mathrm{PNF}_1(w), w'' = \mathrm{PNF}_0(w)$. Then $P_{w'}(n) = F_w^1(n)$ and $P_{w''}(n) = n - F_w^0(n) = f_w^1(n)$.*

Lemma 8. *Let $w \in \{0, 1\}^\omega$. Then $\mathrm{PNF}_1(w)$ is the unique 1-prefix normal word w' s.t. $F_{w'}^1 = F_w^1$. Similarly, $\mathrm{PNF}_0(w)$ is the unique 0-prefix normal word w'' s.t. $F_{w''}^0 = F_w^0$.*

Example 3. For the two prefix normal forms and the maximum-1s and maximum-0s functions of the Fibonacci word $\mathsf{f} = 01001010010010100101\cdots$, see Table 1.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$F_{\mathsf{f}}^0(n)$	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	10	11	12	12	13
$F_{\mathsf{f}}^1(n)$	1	1	2	2	2	3	3	4	4	4	5	5	5	6	6	7	7	7	8	8
$\mathrm{PNF}_0(\mathsf{f})$	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0	1	0
$\mathrm{PNF}_1(\mathsf{f})$	1	0	1	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0	1	0

Table 1. The maximum number of 0s and 1s ($F_{\mathsf{f}}^0(n)$ and $F_{\mathsf{f}}^1(n)$ resp.) for all $n = 1, \ldots, 20$ of the Fibonacci word f, and the prefix normal forms of f.
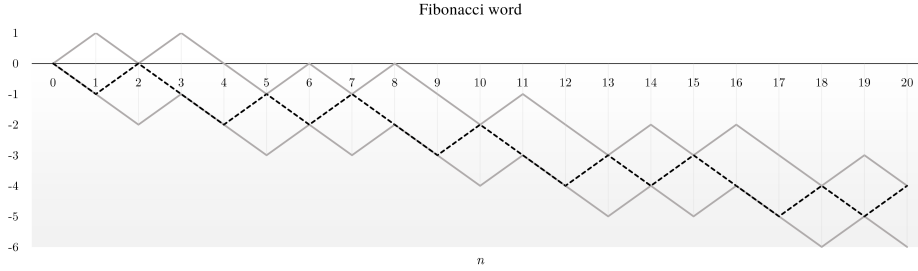
Fig. 1. The Fibonacci word (dashed) and its prefix normal forms (solid). A 1 corresponds to a diagonal segment in direction NE, a 0 in direction SE. On the x-axis the length of the prefix, on the y-axis, number of 1s minus number of 0s in the prefix.

Now we can connect the prefix normal forms of w to the abelian complexity of w in the following way. Given $w' = \mathrm{PNF}_1(w)$ and $w'' = \mathrm{PNF}_0(w)$, the number of Parikh vectors of k-length factors is precisely the difference in 1s in the prefix of length k of w' and of w'' plus 1. For example, Fig. 1 shows the prefix normal forms of the Fibonacci word. The vertical line at 5 cuts through points $(5, -1)$ and $(5, -3)$, meaning that there are two Parikh vectors of factors of length 5, namely $(2, 3)$ and $(1, 4)$. The Fibonacci word, being a Sturmian word, has constant abelian complexity 2. An example with unbounded abelian complexity is the Champernowne word, whose prefix normal forms are 1^ω resp. 0^ω (Fig. 4, Appx.).

Theorem 3. *Let $w, v \in \{0, 1\}^\omega$.*

1. $\psi_w(n) = P_{w'}(n) - P_{w''}(n) + 1$, *where $w' = \mathrm{PNF}_1(w)$ and $w'' = \mathrm{PNF}_0(w)$.*
2. $\Pi(w) = \Pi(v)$ *if and only if* $\mathrm{PNF}_0(w) = \mathrm{PNF}_0(v)$ *and* $\mathrm{PNF}_1(w) = \mathrm{PNF}_1(v)$.

Theorem 3 means that if we know the prefix normal forms of a word, then we can compute its abelian complexity. Conversely, the abelian complexity is the *width* of the area enclosed by the two words $\mathrm{PNF}_1(w)$ and $\mathrm{PNF}_0(w)$. In general, this fact alone does not give us the PNFs; but if we know more about the word itself, then we may be able to compute the prefix normal forms, as we will see in the case of the paperfolding word. We will now give two examples of the close connection between abelian complexity and prefix normal forms, using some recent results about the abelian complexity of infinite words.

1. The paperfolding word. The first few characters of the ordinary paperfolding word are given by

$$\mathfrak{p} = 0010011000110110001001110011011\cdots$$

The paperfolding word was originally introduced in [11]. One definition is given by: $\mathfrak{p}_n = 0$ if $n' \equiv 1 \bmod 4$ and $\mathfrak{p}_n = 1$ if $n' \equiv 3 \bmod 4$, where n' is the unique odd integer such that $n = n'2^k$ for some k [16]. The abelian complexity

function of the paperfolding word was fully determined in [16], giving the following initial values of $\psi_{\mathfrak{p}}(n)$, for $n \geq 1$: $2, 3, 4, 3, 4, 5, 4, 3, 4, 5, 6, 5, 4, 5, 4, 3, 4, 5, 6, 5$, and a recursive formula for all values. The authors note that for the paperfolding word, it holds that if $u \in Fct(\mathfrak{p})$, then also $\overline{u^{\mathrm{rev}}} \in Fct(\mathfrak{p})$. This implies

$$F_{\mathfrak{p}}^1(n) = F_{\mathfrak{p}}^0(n) \text{ for all } n, \text{ and thus } \mathrm{PNF}_0(\mathfrak{p}) = \overline{\mathrm{PNF}_1(\mathfrak{p})}.$$

Moreover, from Thm. 3 we get that $F_{\mathfrak{p}}^1(n) = P_{\mathrm{PNF}_1(\mathfrak{p})}(n) = (\psi_{\mathfrak{p}}(n) + n - 1)/2$, and thus we can determine the prefix normal forms of $\mathfrak{p}$ as shown in Fig. 2. This same argument holds for all words with a symmetric property similar to the paperfolding word:

Lemma 9. *Let $w \in \{0, 1\}^\omega$. If for all $u \in Fct(w)$, it holds that $\overline{u} \in Fct(w)$ or $\overline{u^{\mathrm{rev}}} \in Fct(w)$, then $F_w^1(n) = F_w^0(n)$ for all $n, \mathrm{PNF}_0(w) = \overline{\mathrm{PNF}_1(w)}$, and $F_w^1(n) = (\psi_w(n) + n - 1)/2$.*
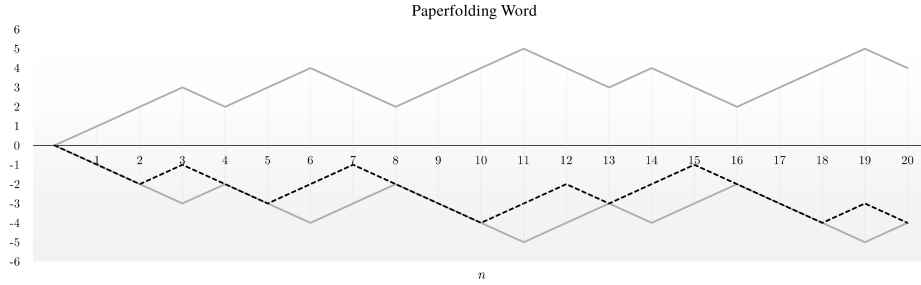

Fig. 2. The paperfolding word (dashed) and its prefix normal forms (solid).

2. Morphic images under the Thue-Morse morphism. The Thue-Morse word beginning with 0, which we denote by **tm**, is one of the two fixpoints of the Thue-Morse morphism μ_{TM}, where $\mu_{\mathrm{TM}}(0) = 01$ and $\mu_{\mathrm{TM}}(1) = 10$:

$$\mathbf{tm} = \mu_{\mathrm{TM}}^{(\omega)}(0) = 0110100110010110100101100110 1001 \cdots$$

The word **tm** has abelian complexity function $\psi_{\mathbf{tm}}(n) = 2$ for n odd and $\psi_{\mathbf{tm}}(n) = 3$ for $n > 1$ even [18]. Since **tm** fulfils the condition that $u \in Fct(\mathbf{tm})$ implies $\overline{u} \in Fct(\mathbf{tm})$, we can apply Lemma 9, and compute the prefix normal forms of **tm** as $\mathrm{PNF}_1(\mathbf{tm}) = 1(10)^\omega$ and $\mathrm{PNF}_0(\mathbf{tm}) = 0(01)^\omega$, see Fig. 3.

For the proof of the abelian complexity of **tm** in [18], the Parikh vectors were computed for each length, so we could have got the prefix normal forms directly (without Lemma 9). Moreover, a much more general result was given in [18]:

Theorem 4 ([18]). *For an aperiodic infinite binary word w, $\psi_w = \psi_{\mathbf{tm}}$ if and only if $w = \mu_{TM}(w')$ or $w = 0\mu_{TM}(w')$ or $w = 1\mu_{TM}(w')$ for some word w'.*
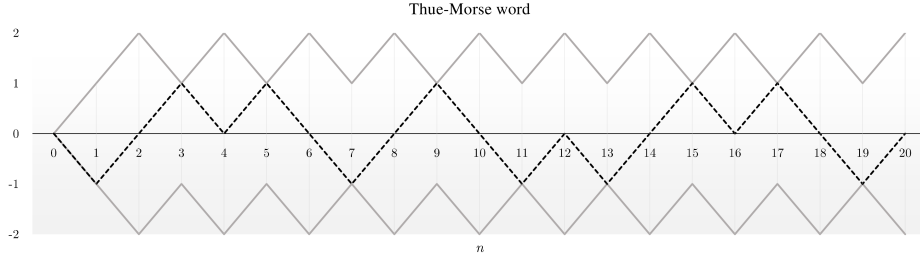
Fig. 3. The Thue-Morse word (dashed) and its prefix normal forms (solid).

The abelian complexity function does not in general determine the prefix normal forms, as can be seen on the example of Sturmian words, which all have the same abelian complexity function but different prefix normal forms. However, $\psi_{\mathbf{tm}}$ does, due to its values $\psi_{\mathbf{tm}}(n) = 2$ for n odd and $= 3$ for n even, and to the fact that both $F^1_{\mathbf{tm}}$ and $F^0_{\mathbf{tm}}$ have difference function with values from $\{0, 1\}$: notice that the only pair of such functions with width 2 resp. 3 are the PNFs of $\mathbf{tm}$. Therefore, we can deduce the following from Theorem 4:

Corollary 3. *For an aperiodic infinite binary word w, $\mathrm{PNF}_1(w) = 1(10)^\omega$ and $\mathrm{PNF}_0 = 0(01)^\omega$ if and only if $w = \mu_{TM}(w')$ or $w = 0\mu_{TM}(w')$ or $w = 1\mu_{TM}(w')$ for some word w'.*

To conclude this section, we return to the question of how many 1s need to be prepended to make the Thue-Morse word prefix normal.

Lemma 10. *We have $11\mathbf{tm} \in \mathcal{L}$. This is minimal since $1\mathbf{tm}$ is not prefix normal.*

5.3 Prefix normal forms of Sturmian words. Let w be a Sturmian word. As we saw in Section 4, the only 1-prefix normal word in the class of Sturmian words with the same slope α is the upper mechanical word $s'_{\alpha,0} = 1c_\alpha$.

Theorem 5. *Let w be an irrational mechanical word with slope α, i.e. a Sturmian word. Then $\mathrm{PNF}_1(w) = 1c_\alpha$ and $\mathrm{PNF}_0(w) = 0c_\alpha$, where c_α is the characteristic word of slope α.*

6   Prefix normal words and lexicographic order

In this section, we study the relationship between lexicographic order and prefix normality. Note that for coherence with the rest of the paper, in the definition of Lyndon words, necklaces, and prenecklaces, we use lexicographically *greater* rather than *smaller*. Clearly, this is equivalent to the usual definitions up to renaming characters.

Thus a finite *Lyndon word* is one which is lexicographically strictly greater than all of its conjugates: w is Lyndon if and only if for all non-empty u, v s.t.

$w = uv$, we have $w >_{\text{lex}} vu$. A *necklace* is a word which is greater than or equal to all its conjugates, and a *prenecklace* is one which can be extended to become a necklace, i.e. which is the prefix of some necklace [15, 19]. As we saw in the introduction, in the finite case, prefix normality and Lyndon property are orthogonal concepts. However, the set of finite prefix normal words is included in the set of prenecklaces [6].

An infinite word is *Lyndon* if an infinite number of its prefixes is Lyndon [24]. In the infinite case, we have a similar situation as in the finite case. There are words which are both Lyndon and prefix normal: $10^\omega, 110(10)^\omega$; Lyndon but not prefix normal: $11100(110)^\omega$; prefix normal but not Lyndon: $(10)^\omega$; and neither of the two: $(01)^\omega$. Next we show that a prefix normal word cannot be lexicographically smaller than any of its suffixes. Let $shift_i(w) = w_i w_{i+1} w_{i+2} \cdots$ denote the infinite word v s.t. $w = w_1 \cdots w_{i-1} v$, i.e. v is w starting at position i.

Lemma 11. *Let $w \in \mathcal{L}_{\text{inf}}$. Then $w \geq_{\text{lex}} shift_i(w)$ for all $i \geq 1$.*

In the finite case, it is easy to see that a word w is a prenecklace if and only if $w \geq_{\text{lex}} v$ for every suffix v of w. This motivates our definition of infinite prenecklaces. The situation is the same as in the finite case: prefix normal words form a proper subset of prenecklaces.

Definition 9. *Let $w \in \{0,1\}^\omega$. Then w is an* infinite prenecklace *if for all $i \geq 1$, $w \geq_{\text{lex}} shift_i(w)$. We denote by $\mathcal{P}_{\text{inf}}$ the set of infinite prenecklaces.*

Proposition 2. *We have $\mathcal{L}_{\text{inf}} \subsetneq \mathcal{P}_{\text{inf}}$.*

Another interesting relationship is that between lexicographic order and the prefix normal forms of an infinite word. In [17], two words were associated to an infinite binary word w, called $\max(w)$ resp. $\min(w)$, defined as the word whose prefix of length n is the lexicographically greatest (resp. smallest) n-length factor of w. It is easy to see that these words always exists. It was shown in [17]:[2]

Theorem 6 ([17]). *Let w be an infinite binary word. Then*

1. *w is (rational or irrational) mechanical with its intercept equal to its slope if and only if $0w \leq_{\text{lex}} \min(w) \leq_{\text{lex}} \max(w) \leq_{\text{lex}} 1w$, and*
2. *w is characteristic Sturmian if and only if $\min(w) = 0w$ and $\max(w) = 1w$.*

Lemma 12. *For $w \in \{0,1\}^\omega$, $\text{PNF}_1(w) \geq_{\text{lex}} \max(w)$ and $\text{PNF}_0(w) \leq_{\text{lex}} \min(w)$.*

From Theorems 5 and 6 we get the following corollary:

Corollary 4. *Let w be an infinite binary word. Then w is characteristic Sturmian if and only if $0w = \text{PNF}_0(w) = \min(w)$ and $1w = \text{PNF}_1(w) = \max(w)$.*

---

[2] Note the different terminology in [17], e.g. characteristic word $\rightarrow$ proper standard Sturmian, see the Appendix for a detailed comparison.

References

1. A. Amir, T. M. Chan, M. Lewenstein, and N. Lewenstein. On hardness of jumbled indexing. In *41st International Colloquium on Automata, Languages, and Programming (ICALP 2014)*, volume 8572 of *LNCS*, pages 114–125, 2014.
2. F. Blanchet-Sadri, N. Fox, and N. Rampersad. On the asymptotic abelian complexity of morphic words. *Advances in Applied Mathematics*, 61:46–84, 2014.
3. A. Blondin Massé, J. de Carufel, A. Goupil, M. Lapointe, É. Nadeau, and É. Vandomme. Leaf realization problem, caterpillar graphs and prefix normal words. *Theoret. Comput. Sci.*, 732:1–13, 2018.
4. P. Burcsi, F. Cicalese, G. Fici, and Zs. Lipták. Algorithms for Jumbled Pattern Matching in Strings. *Int. J. of Found. Comput. Sci.*, 23:357–374, 2012.
5. P. Burcsi, G. Fici, Zs. Lipták, F. Ruskey, and J. Sawada. On combinatorial generation of prefix normal words. In *Proc. of the 25th Ann. Symp. on Comb. Pattern Matching (CPM 2014)*, volume 8486 of *LNCS*, pages 60–69, 2014.
6. P. Burcsi, G. Fici, Zs. Lipták, F. Ruskey, and J. Sawada. On prefix normal words and prefix normal forms. *Theoret. Comput. Sci.*, 659:1–13, 2017.
7. J. Cassaigne and I. Kaboré. Abelian complexity and frequencies of letters in infinite words. *Int. Journal of Found. Comput. Sci.*, 27(05):631–649, 2016.
8. T. M. Chan and M. Lewenstein. Clustered integer 3SUM via additive combinatorics. In *Proc. of the 47th Ann. ACM on Symp. on Theory of Computing (STOC 2015)*, pages 31–40, 2015.
9. F. Cicalese, Zs. Lipták, and M. Rossi. Bubble-Flip - A new generation algorithm for prefix normal words. *Theor. Comput. Sci.*, 743:38–52, 2018.
10. L. F. I. Cunha, S. Dantas, T. Gagie, R. Wittler, L. A. B. Kowada, and J. Stoye. Fast and Simple Jumbled Indexing for Binary Run-Length Encoded Strings. In *28th Annual Symposium on Combinatorial Pattern Matching (CPM 2017)*, volume 78 of *LIPIcs*, pages 19:1–19:9, 2017.
11. C. Davis and D. Knuth. Number representations and dragon curves, I, II. *J. Recr. Math.*, 3:133–149 and 161–181, 1970.
12. G. Fici and Zs. Lipták. On prefix normal words. In *Proc. of the 15th Intern. Conf. on Developments in Language Theory (DLT 2011)*, volume 6795 of *LNCS*, pages 228–238, 2011.
13. T. Gagie, D. Hermelin, G. M. Landau, and O. Weimann. Binary jumbled pattern matching on trees and tree-like structures. *Algorithmica*, 73(3):571–588, 2015.
14. I. Kaboré and B. Kientéga. Abelian complexity of Thue-Morse word over a ternary alphabet. In *Proc. of the 11th Int. Conf. on Combinatorics on Words WORDS 2017*, volume 10432 of *LNCS*, pages 132–143. Springer, 2017.
15. M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge Univ. Press, 2002.
16. B. Madill and N. Rampersad. The abelian complexity of the paperfolding word. *Discrete Mathematics*, 313(7):831–838, 2013.
17. G. Pirillo. Inequalities characterizing standard Sturmian and episturmian words. *Theor. Comput. Sci.*, 341(1-3):276–292, 2005.
18. G. Richomme, K. Saari, and L. Q. Zamboni. Abelian complexity of minimal subshifts. *J. London Math. Society*, 83(1):79–95, 2011.
19. F. Ruskey, C. Savage, and T. Wang. Generating necklaces. *J. Algorithms*, 13(3):414–430, 1992.
20. F. Ruskey, J. Sawada, and A. Williams. Binary bubble languages and cool-lex order. *J. Comb. Theory, Ser. A*, 119(1):155–169, 2012.

21. J. Sawada and A. Williams. Efficient oracles for generating binary bubble languages. *Electr. J. Comb.*, 19(1):P42, 2012.
22. J. Sawada, A. Williams, and D. Wong. Inside the Binary Reflected Gray Code: Flip-Swap languages in 2-Gray code order. Unpublished manuscript, 2017.
23. N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences. http://oeis.org.
24. R. Siromoney, L. Mathew, V. Dare, and K. Subramanian. Infinite Lyndon words. *Inf. Proc. Letters*, 50:101–104, 1994.
25. O. Turek. Abelian complexity function of the Tribonacci word. *J. of Integer Sequences*, 18, Article 15.3.4, 2015.

Appendix

A1: Missing proofs

Proof (Lemma 2). Let us write $v = ux^\omega$ and x not a suffix of u.

For $i = 0, 1, \ldots, |x| - 1$, let y_i be the prefix of length $|u| + i$ of v, i.e., $y_i = ux_1x_2 \cdots x_i$. Trivially, if for some i we have that $\delta(y_i) \leq \delta(v)$ the claim directly follows from y_i being a finite prefix of v.

Let us now assume that for each $i = 0, 1, \ldots |x| - 1$ it holds that $\delta(v) < \delta(y_i)$ and let $i^* = \min\{i \mid \delta(y_i) \leq \delta(y_j) \text{ for each } j \neq i\}$, hence $\delta(v) < \delta(y_{i^*})$.

For every $n \geq |u| + |x|$ let $i_n = |u| + ((n - |u|) \bmod |x|)$ and $k_n = \lfloor (n - |u|)/|x| \rfloor$, i.e., $|u| \leq i_n \leq |u| + |x| - 1$ and $n = i_n + k_n|x|$.

Then, we have that

$$D_v(n) = \frac{|y_{i_n}|_1 + k_n|x|_1}{|y_{i_n}| + k_n|x|} \geq \min\{\delta(y_{i_n}), \delta(x)\} \geq \min\{\delta(y_{i^*}), \delta(x)\}. \qquad (1)$$

Moreover, we also have that

$$\lim_{k \to \infty} D_v(|u| + i^* + k|x|) = \lim_{k \to \infty} \frac{|y_{i^*}|_1 + k|x|_1}{|y_{i^*}| + k|x|} = \delta(x). \qquad (2)$$

We cannot have $\delta(x) \geq \delta(y_{i^*})$, since by (1) $\delta(y_{i^*})$ is a rational lower bound on $D_v(n)$ (for each $n \geq 1$) which is achieved by $D_v(|u| + i^*)$, contradicting the standing hypothesis $\delta(v) < \delta(y_{i^*})$.

Therefore, we must have $\delta(x) < \delta(y_{i^*})$, and from (1) we have $D_v(n) \geq \delta(x)$ and from (2) we also have that for each $\varepsilon > 0$ there exists $k > 0$ such that $D_v(|u| + i^* + k|x|) < \delta(x) + \varepsilon$. Therefore, $\delta(v) = \inf\{D_v(n) \mid n \geq 1\} = \delta(x)$, which is a rational number, since x is a finite string. $\qquad \square$

Proof (Lemma 3). The statement is a direct consequence of the following claim.
Claim The following properties hold

1. $\delta(v^{(i)}) \geq a_i$ for each $i \geq 1$;
2. $\iota(v^{(i)}) = |v^{(i)}|$ for each $i \geq 1$;
3. $\delta(v^{(i)}) < \delta(v^{(i-1)})$ for each $i \geq 2$;
4. $|v^{(i)}|_1 > |v^{(i-1)}|_1$ for each $i \geq 2$;
5. $\delta(v^{(i)}) \leq a_i \left(\frac{k_i|v^{(i-1)}|_1}{k_i|v^{(i-1)}|_1 - a_i} \right)$ for each $i \geq 2$.

Proof of the claim. By direct inspection we have that properties 1 and 2 hold for $v^{(1)}$. We now argue by induction. Fix $i > 1$ and let us assume that properties 1 and 2 hold for $v^{(i-1)}$. Then, since $a_i < a_{i-1}$ we have

$$\frac{|v^{(i-1)}|_1}{a_i} > \frac{|v^{(i-1)}|_1}{a_{i-1}} \geq |v^{(i-1)}|,$$

where the last inequality follows from property 1. Therefore, $\left(\frac{|v^{(i-1)}|_1 - a_i|v^{(i-1)}|}{a_i} \right) > 0$, hence there exists $k_i > 1$ such that $\left\lfloor k_i \left(\frac{|v^{(i-1)}|_1 - a_i|v^{(i-1)}|}{a_i} \right) \right\rfloor > \ell_{i-1}$. In particular, ℓ_i is well defined.

By property 2, we have $\iota(v^{(i-1)}) = |v^{(i-1)}|$ hence by Proposition 1, we have $D_{\text{flipext}^\omega(v^{(i-1)})}(k|v^{(i-1)}|) = \delta(v^{(i-1)})$ and also $\delta(\text{pref}_{\text{flipext}^\omega(v^{(i-1)})}(k_i|v^{(i-1)}|)) = \delta(v^{(i-1)})$.

Moreover, since $\ell_i > 0$ it is not hard to see from the definition of $v^{(i)}$ that

$$\delta(v^{(i)}) = D_{v^{(i)}}(|v^{(i)}|) = \frac{k_i|v^{(i-1)}|_1}{k_i|v^{(i-1)}| + \ell_i} < \delta(v^{(i-1)}), \tag{3}$$

which shows that property 3 and property 2 hold for $v^{(i)}$. In addition, because of $k_i > 1$ and (by Proposition 1)

$$|v^{(i)}|_1 = |\text{pref}_{\text{flipext}^\omega(v^{(i-1)})}(k_i|v^{(i-1)}|)|_1 = k_1|v^{(i-1)}|_1$$

it follows that property 4 also holds for $v^{(i)}$.

The definition of ℓ_i together with the well known property $x - 1 < \lfloor x \rfloor \leq x$ imply that

$$\frac{k_i}{a_i}\left(|v^{(i-1)}|_1 - a_i|v^{(i-1)}|\right) - 1 < \ell_i \leq k_i\left(\frac{|v^{(i-1)}|_1}{a_i} - |v^{(i-1)}|\right). \tag{4}$$

Using the right inequality of (4) in (3) we have $\delta(v^{(i)}) \geq a_i$ showing that property 1 holds for $v^{(i)}$.

In addition, using the left inequality of (4) in (3) we have

$$\delta(v^{(i)}) \leq a_i\left(\frac{k_i|v^{(i-1)}|_1}{k_i|v^{(i-1)}|_1 - a_i}\right)$$

showing that property 5 holds for $v^{(i)}$. The proof of the claim is complete.

In order to see that v is aperiodic, it is enough to observe that $v \neq 0^\omega$ and for each $i \geq 1$ it contains a distinct run of ℓ_i 0s, with ℓ_i being a strictly increasing sequence.

In order to show that $\delta(v) = \alpha$, we will prove that $\lim_{i \to \infty} \delta(v^{(i)}) = \alpha$.

Since, $\lim_{i \to \infty} a_i = \alpha$ and for each $i \geq 1$, $k_i > 1$ and $|v^{(i)}|_1 > |v^{(i-1)}|_1$, we have that

$$\lim_{i \to \infty} a_i \frac{k_i|v^{(i-1)}|_1}{k_i|v^{(i-1)}|_1 - a_i} = \lim_{i \to \infty} a_i = \alpha.$$

Hence, from properties 4 and 5 of the Claim above, we have the desired result $\lim_{i \to \infty} \delta(v^{(i)}) = \lim_{i \to \infty} a_i = \alpha$.

Proof (Lemma 4). First note that $\delta(v) \geq \alpha$ by definition. Now write $v = w0^k1$, and let $u = \text{flipext}(w) = w0^\ell1$. Recall that $\ell = \min\{j \mid w0^j1 \in \mathcal{L}\}$. If $k < \ell$, this implies $\delta(u) < \alpha$, in contradiction to Proposition 1, since $\delta(u) = \delta(w) \geq \alpha$. Thus $k \geq \ell$, from which follows that $v \in \mathcal{L}$.

Proof (Corollary 1). That v is prefix normal follows from Lemma 1 and from Lemma 4, which also implies that $\delta(v) \geq \alpha$. If $\delta(v) > \alpha$ was true, then for a suitably long prefix, we would get a contradiction to the definition of the lazy-α-flipext operation. □

Proof (Lemma 5). Let s_i and v_i denote the ith character of s and v respectively. We argue by induction on i that $v_i = s_i$. The claim is true for $i = 1$ since, directly from the definitions we have $v_1 = 1 = s_1$. Let $n > 1$ and assume that for each $i < n$ we have $v_i = s_i$. For the induction step we argue according to the character s_n.

(i) If $s_n = 1$, by definition $\lceil n\alpha \rceil - \lceil (n-1)\alpha \rceil = 1$. Thus, $\lceil (n-1)\alpha \rceil < n\alpha$. Using this inequality and the induction hypothesis together with the definition of $s'_{\alpha,0}$ we have that $|v_1 \cdots v_{n-1}|_1 = |s_1 \cdots s_{n-1}|_1 = \lceil (n-1)\alpha \rceil < \alpha n$. Therefore $|v_1 \cdots v_{n-1}0|_1 = |v_1 \cdots v_{n-1}|_1 < \alpha n$ which means that $\delta(v_1 \cdots v_{n-1}0) < \alpha$, hence by definition lazy-α-flipext$(v_1 \cdots v_{n-1}) = v_1 \cdots v_{n-1}1$, i.e., $v_n = 1 = s_n$.

(ii) If $s_n = 0$, by definition $\lceil n\alpha \rceil - \lceil (n-1)\alpha \rceil = 0$. Thus, $\lceil (n-1)\alpha \rceil \geq n\alpha$. Using this inequality and the induction hypothesis together with the definition of $s'_{\alpha,0}$ we have that $|v_1 \cdots v_{n-1}| = |s_1 \cdots s_{n-1}| = \lceil (n-1)\alpha \rceil \geq \alpha n$. Therefore $|v_1 \cdots v_{n-1}0|_1 = |v_1 \cdots v_{n-1}|_1 \geq \alpha n$ which means that $\delta(v_1 \cdots v_{n-1}0) \geq \alpha$, hence by definition lazy-α-flipext$(v_1 \cdots v_{n-1}) = v_1 \cdots v_{n-1}0 \cdots 01$, i.e., $v_n = 0 = s_n$. □

Proof (Theorem 2). By definition, α is irrational. Let $s = s'_{\alpha,0}$. Then s is Sturmian and prefix normal by Corollary 2. Let t be a Sturmian word with the same slope α which is also prefix normal. By Fact 1, s and t have the same factors.

Assume, by contradiction, that $s \neq t$, hence there exists $i \geq 1$ such that $|s_1 \cdots s_i|_1 \neq |t_1 \cdots t_i|_1$. Assume, without loss of generality (since we can, if necessary, swap s and t in the following argument), that $|s_1 \cdots s_i|_1 > |t_1 \cdots t_i|_1$. Then, since $s_1 \cdots s_i$ is also a factor of t, there is a $j \geq 1$ such that $t_{j+1} \cdots t_{j+i} = s_1 \cdots s_i$, hence $|t_{j+1} \cdots t_{j+i}|_1 > |t_1 \cdots t_i|_1$ contradicting the assumption that t is prefix normal. □

Proof (Lemma 6). 1. Let $0w$ be Sturmian and let u be some factor of $1w$. If u is a prefix of $1w$, there is nothing to show, therefore let $u \in Fct(w)$, with $|u| = n$ and $|u|_1 = k$. Since $0w$ is Sturmian, we have that the prefix of $0w$ of length n has at least $k - 1$ 1s, thus $P_{1w}(n) \geq k = |u|_1$, as desired. Conversely, if $0w$ is not Sturmian, this means that it is not balanced, therefore there exists a factor u of w s.t. $||u|_1 - |0w_1 \cdots w_{n-1}|_1| \geq 2$, where $|u| = n$. Since w is Sturmian, we have that $||w_1 \cdots w_{n-1}|_1 - |u_1 \cdots u_{n-1}|_1| \leq 1$ and $||w_1 \cdots w_{n-1}|_1 - |u_2 \cdots u_n|_1| \leq 1$. Let $|w_1 \cdots w_{n-1}|_1 = k$, then this implies, by a case-by-case consideration, that $|u_1 \cdots u_{n-1}|_1 = |u_2 \cdots u_n|_1 = k + 1$, and thus $|1w_1 \cdots w_{n-1}|_1 = k + 1 < k + 2 = |u|_1$, showing that $1w$ is not prefix normal.

2. First note that a Sturmian word of slope α cannot have a run of 1s of length $1/(1 - \alpha)$. To see this, it is enough to argue about the upper mechanical word of slope α and intercept 0 (since all the other words with the same slope have the same set of factors). Let us write $s = s_{\alpha,0} = s_1 s_2 \cdots$

Now s has a run of n 1s iff there exists an $i \geq 0$ such that $s_{i+1} = s_{i+2} = \cdots = s_{i+n} = 1$. By the definition of mechanical words, we have that the last condition is equivalent to

$$\lceil \alpha(i+n) \rceil - \lceil \alpha i \rceil = n.$$

On the other hand, if $n \geq \frac{1}{1-\alpha}$, i.e., $\alpha \leq \frac{n-1}{n}$ we have that the sum of the character $\sum_{j=1}^{n} s_{i+j}$ satisfies

$$\begin{aligned}
\sum_{j=1}^{n} s_{i+j} &= \lceil \alpha(i+n) \rceil - \lceil \alpha i \rceil \\
&\leq \lceil \alpha i \rceil + \lceil \alpha n \rceil - \lceil \alpha i \rceil \\
&= \lceil \alpha n \rceil \\
&< \alpha n + 1 \\
&\leq \frac{n-1}{n} \times n + 1 = n.
\end{aligned}$$

i.e., strictly smaller than n, i.e., we have a contradiction $s_{i+1} \cdots s_{i+n} \neq 1^n$.

Now fix $n = \lceil 1/(1-\alpha) \rceil$ and let $w' = 1^n w$. Let $u \in Fct(w)$. Since, as shown above, 1^n is not a factor, if $|u| \leq n$, there is nothing to show. So let $|u| = n + m$. Then $|u_1 \cdots u_n|_1 \leq n - 1$, and since w is balanced, we have that $|w_1 \cdots w_m|_1 \geq |u_{n+1} \cdots u_{n+m}|_1 - 1$, yielding that $P_{w'}(n+m) \geq n + |u_{n+1} \cdots u_{n+m}|_1 - 1 \geq |u|_1$. $\qquad \square$

Proof (Lemma 7). We are going to show that every factor u of z satisfies the prefix normal condition $|u|_1 \leq P_z(|u|)$. It is not hard to see that we can limit ourselves to only considering factors u such that u does not overlap with the prefix of z of the same length.

If $|u| \leq nc$ then $|u|_1 \leq |u| = P_z(|u|)$. Assume now that $u = u'u''$ with $|u'| = nc$ and $|u''| > 0$. Since u' is a factor of w of size nc the condition that w does not contain a factor 1^n implies that u' contains at least c 0s, i.e., $|u'|_1 \leq |u'| - c$. Moreover, since w is c-balanced, we have that $|u''|_1 \leq P_w(|u''|) + c$. Therefore, observing that $\text{pref}_z(|u|) = \text{pref}_z(|u'| + |u''|) = 1^{nc} \text{pref}_w(|u''|)$ we have that $P_z(|u|) = nc + P_w(|u''|) \geq |u'|_1 + |u''|_1 = |u|_1$. $\qquad \square$

Proof (Lemma 8). Let $w' = \text{PNF}_1(w)$ and $w'' = \text{PNF}_0(w)$. First note that, by construction, $F_{w'}^1 = F_w^1$ and $F_{w''}^0 = F_w^0$. It is easy to see that w' is 1-prefix normal and w'' is 0-prefix normal. For uniqueness, note that for $a \in \{0,1\}$ and an a-prefix normal word v, we have $\text{PNF}_a(v) = v$. $\qquad \square$

Proof (Theorem 3).

1. Fix an integer $n \geq 1$. By definition, we have that for every factor u of w of length n we have $n - F_w^0(n) \leq |u|_1 \leq F_w^1(n)$. Therefore $\psi_w(n) \leq F_w^1(n) - (n - F_w^0(n)) + 1$.

Conversely, since w contains a factor u' of length n with $F_w^1(n)$ many 1s and a factor u'' of length n with $n - F_w^0(n)$ many 1s, if we scan w between an occurrence

of u' and an occurrence of u'', for each $x \in \{|u''|_1, \ldots, |u'|_1\}$ there must be a factor u''' of size n such that $|u'''|_1 = x$. Therefore $\psi_w(n) \geq F_w^1(n) - (n - F_w^0(n)) + 1$. We can conclude that $\psi_w(n) = F_w^1(n) - (n - F_w^0(n)) + 1$. The desired result then follows by observing that $n - F_w^0(n) = n - |\operatorname{pref}_{\mathrm{PNF}_0(w)}(n)|_0 = P_{\mathrm{PNF}_0(w)}(n)$ and $F_w^1(n) = P_{\mathrm{PNF}_1(w)}(n)$.

2. Follows directly from Observation 1. □

Proof (Lemma 9). Same as for the special case of the paperfolding word. □

Proof (Lemma 10). We will show that for every prefix, the number of 1s in the prefix of $11\mathbf{tm}$ is greater than or equal to the the number of 1s in the prefix of $\mathrm{PNF}_1(\mathbf{tm})$ of the same length. Let $v = \mathrm{PNF}_1(\mathbf{tm})$ and $u = 11\mathbf{tm}$. It is easy to see that $P_v(n) = \lfloor \frac{n}{2} \rfloor + 1$ and

$$P_u(n) = \begin{cases} \frac{n}{2} + 1 & \text{if } n \text{ is even} \\ \lfloor \frac{n}{2} \rfloor + 2 & \text{if } n \text{ is odd and } u_n = 1 \\ \lfloor \frac{n}{2} \rfloor + 1 & \text{if } n \text{ is odd and } u_n = 0 \end{cases}$$

Thus for all $n \geq 1$ it holds that $P_u(n) \geq P_v(n)$, implying that $11\mathbf{tm} \in \mathcal{L}$.

For minimality, note that $1\mathbf{tm}$ is not prefix normal, since 11 is a factor of $\mathbf{tm}$. □

Proof (Theorem 5). Since the characteristic word c_α has the same slope as w, we have $Fct(w) = Fct(c_\alpha)$ by Fact 1. The abelian complexity of w is constant 2 [18], thus a factor of length k can have either $F_w^1(k)$ or $F_w^1(k) - 1$ 1s. Let us call a factor u of w *heavy* if $|u|_1 = F_w^1(k)$, and *light* otherwise. We have to show that every prefix of $1c_\alpha$ is heavy. It is known [15] that the prefixes of the characteristic word are precisely the reverses of its right special factors, where a factor u is called right special if both $u0$ and $u1$ are factors. Thus, every prefix v of $1c_\alpha$ has the form $v = 1u^{\mathrm{rev}}$, where both $u1$ and $u0$ are factors of w, therefore $v = 1u^{\mathrm{rev}}$ is heavy. The fact that $\mathrm{PNF}_0(w) = 0c_\alpha$ follows analogously. □

Proof (Lemma 11). Assume that there exists a suffix $v = shift_i(w)$ of w s.t. $v >_{\mathrm{lex}} w$. Then there is an index j with $v_1 \cdots v_{j-1} = w_1 \cdots w_{j-1}$ and $v_j > w_j$, implying $v_j = 1$ and $w_j = 0$. But then $|w_i \cdots w_{i+j-1}|_1 = |v_1 \cdots v_j|_1 > |w_1 \cdots w_j|_1$, in contradiction to $w \in \mathcal{L}_{\mathrm{inf}}$. □

Proof (Proposition 2). The inclusion follows from Lemma 11. An example of a word which is an infinite prenecklace but not prefix normal is $11100(110)^\omega$. □

Proof (Lemma 12). Assume otherwise, and let $w' := \mathrm{PNF}_1(w), v := \max(w)$. If $w' < v$, then there is an index j s.t. $w'_1 \cdots w'_{j-1} = v_1 \cdots v_{j-1}$ and $w'_j = 0$ and $v_j = 1$. This implies that $v_1 \cdots v_j$ has one more 1s than $w'_1 \cdots w'_j$. But $|w'_1 \cdots w'_j|_1 = F_w^1(j)$, a contradiction, since $v_1 \cdots v_j$ is a factor of w. The second claim follows analogously. □

A2: A note on terminology

The terminology in [17] differs from ours (we are following [15]). In order to help the reader to reference correctly the results we want to highlight the differences. (*i*) a periodic Sturmian in [17] is a rational mechanical word, (*ii*) a proper Sturmian word in [17] is an irrational mechanical word (i.e., a Sturmian word), and (*iii*) a standard Sturmian word in [17] is a mechanical word for with intercept $\tau = \alpha$, thus a proper standard Sturmian word is a characteristic Sturmian word c_α. Note that all mechanical words in [17] are defined for $n \geq 1$ since the definition of mechanical word is: the lower mechanical word is defined as $s_{\alpha,\tau}(n) = \lfloor \alpha(n+1) + \tau \rfloor - \lfloor \alpha n + \tau \rfloor$ for $n \geq 1$, and analogously for the upper mechanical word. Therefore, an intercept $\tau = 0$ in [17] is equivalent to an intercept of $\tau = \alpha$ (the slope) in [15].
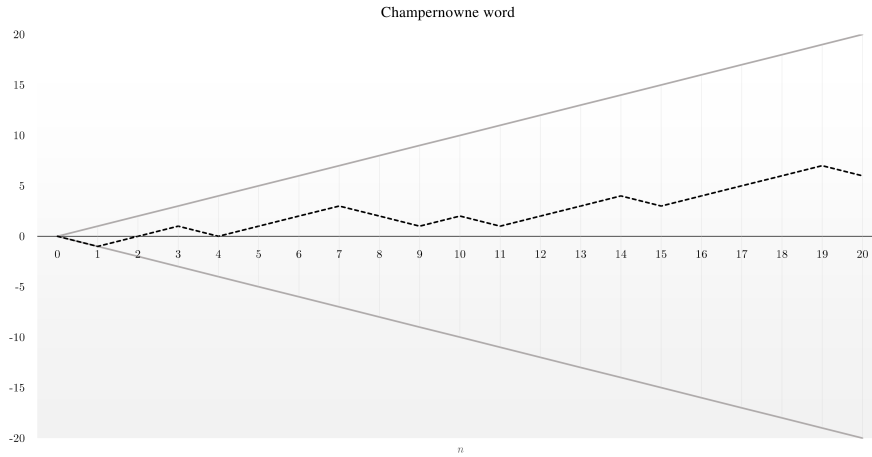
A3: Additional Figures

Fig. 4. The Champernowne word (dashed) and its prefix normal forms (solid).