Acta Polytechnica **65**(5):534–538, 2025
https://doi.org/10.14311/AP.2025.65.0534

Published by the Czech Technical University in Prague

FREQUENCIES OF LETTERS IN INFINITE k-BALANCED SEQUENCES

Ľubomíra Dvořáková*, Edita Pelantová

Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Department of Mathematics, Trojanova 13, 120 00 Prague, Czech Republic

* corresponding author: lubomira.dvorakova@fjfi.cvut.cz

Abstract. The frequency of letters in a symbolic sequence $\mathbf{u}$ over a finite alphabet is one of the basic characteristics of $\mathbf{u}$. The notion of k-balancedness captures the property that the number of any letter occurring in two arbitrary factors of $\mathbf{u}$ of equal length differs at most by k. For a fixed integer k and alphabet size $d \in \mathbb{N}$, we discuss possible frequencies of letters in k-balanced d-ary sequences. For the size d of the alphabet, we introduce the notion of balancedness threshold $BT(d)$ and provide an upper bound on it, where $BT(d)$ is the minimum k such that there exists a k-balanced sequence over a d-letter alphabet for all possible letter frequencies.

Keywords: Letter frequency, balanced sequences, balancedness threshold, factor complexity.

1. Introduction

This paper is devoted to studying the relation between frequencies of letters and k-balancedness in sequences (also called infinite words). Let us first introduce these notions. Consider a sequence $\mathbf{u} = u_0u_1u_2\cdots$ of symbols from a finite alphabet $\mathcal{A} = \{1, 2, \ldots, d\}$. The *frequency of a letter* a in $\mathbf{u}$ is the limit (if it exists):

$$f_a = \lim_{n\to\infty} \frac{\#\{i < n : u_i = a\}}{n}.$$

If every letter $a \in \mathcal{A}$ has a well-defined frequency in $\mathbf{u}$, then $\sum_{a\in\mathcal{A}} f_a = 1$. The vector $\vec{f}_{\mathbf{u}} = (f_a)_{a\in\mathcal{A}}$ is called the *frequency vector* of $\mathbf{u}$.

A *word* $w = w_0w_1\cdots w_{n-1}$ over $\mathcal{A}$ is a finite sequence of letters w_i from $\mathcal{A}$. Its *length* $|w|$ equals n. To denote the number of occurrences of a letter a in w, we use $|w|_a$. The set of all words over the alphabet $\mathcal{A}$ (including the empty word) is denoted $\mathcal{A}^*$. The word w is a *factor* of $\mathbf{u} = u_0u_1u_2\cdots$ if there exists $i \in \mathbb{N}$ such that $w = u_iu_{i+1}\cdots u_{i+|w|-1}$. We say that the sequence $\mathbf{u}$ is recurrent if every factor of $\mathbf{u}$ occurs in $\mathbf{u}$ infinitely many times. The *language* $\mathcal{L}(\mathbf{u})$ of a sequence $\mathbf{u}$ is the set of factors occurring in $\mathbf{u}$. The *factor complexity* of a sequence $\mathbf{u}$ is the mapping $\mathcal{C} : \mathbb{N} \to \mathbb{N}$, where:

$$\mathcal{C}(n) = \#\{w \in \mathcal{L}(\mathbf{u}) \ : \ |w| = n\}.$$

We say that $\mathbf{u}$ is *k-balanced* if for any two factors v, w of $\mathbf{u}$ of the same length and for any letter $a \in \mathcal{A}$ holds $\big||v|_a - |w|_a\big| \le k$. Obviously, a k-balanced sequence is K-balanced for any $K > k$.

The study of 1-balanced sequences over a binary alphabet $\{a, b\}$ was initiated by Hedlund and Morse [1]. They showed that 1-balancedness requires some particular properties of the sequence. If a 1-balanced sequence is eventually periodic, then f_a and f_b are rational. In the opposite case, a 1-balanced sequence is called *Sturmian*. Hedlund and Morse proved that for each positive vector (f_a, f_b), where $f_a + f_b = 1$, there exists a 1-balanced sequence with such letter frequencies. The class of Sturmian sequences is the most studied class of sequences and there exist a lot of equivalent definitions of Sturmian sequences, see [2–4]. These equivalent definitions allow Sturmian sequences to be generalised to larger alphabets in many different ways, see [2]. On the one hand, one of the most usual generalisations are the Arnoux-Rauzy sequences, however, it is known that the letter frequencies of any Arnoux-Rauzy sequence belong to the Rauzy gasket [5], which is a fractal set of Lebesgue measure zero. On the other hand, for any given letter frequencies, one can construct a sequence of sublinear factor complexity by coding a d-interval exchange transformation. It is, however, known [6] that such a generalisation of Sturmian sequences to d-ary alphabet is almost always unbalanced.

1-balanced sequences over alphabets of size d were described by Hubert [7]. The description implies that for $d \ge 3$, the frequency vector $(f_a)_{a\in\mathcal{A}}$ of 1-balanced sequences takes only a specific form, see Lemma 11. This fact motivates our definition of balancedness threshold for an alphabet of size d.

Definition 1. We say that $k \in \mathbb{N}$ is *frequency restrictive* for d if there exists a positive vector $\vec{f} = (f_1, f_2, \ldots, f_d)$ with $\sum_{i\in\mathcal{A}} f_i = 1$ such that no k-balanced d-ary sequence has the frequency vector $\vec{f}$. *Balancedness threshold* $BT(d)$ is the minimum $k \in \mathbb{N}$ such that k is not frequency restrictive for d.

Obviously, $BT(1) = 0$. It follows from the result by Hedlund and Morse that $BT(2) = 1$. In Lemma 11, we explain why $BT(d) \ge 2$ for every $d \ge 3$. In [8], the sequence coding rectangle exchange transformation is used to prove $BT(3) = 2$. Moreover, the factor complexity of such ternary sequences satisfies $\mathcal{C}(n) \le$

$\alpha n^2\big(1+o(1)\big)$, for parameter $\alpha \in (0,1)$. The equality $BT(3) = 2$ follows also from the properties of ternary hypercubic billiard sequences. This class contains sequences of any given letter frequencies. They are 2-balanced [9] and under an additional condition on momentum, their factor complexity equals n^2+n+1, see [10–12].

The main contribution of this paper is the upper bound on the balancedness threshold.

Theorem 2. *Let d be a positive integer. Then $BT(d) \le \lceil \log_2 d \rceil$.*

On top of it, we discuss factor complexity of d-ary sequences used in the proof of the above theorem. Getting a lower bound on $BT(d)$ is beyond our means. It is unclear whether, for some $d \in \mathbb{N}$, $BT(d) \ge 3$ holds. The characterisation of 2-balanced sequences, which would be helpful for this purpose, is still missing.

2. Frequencies and balancedness in fixed points of morphism

In combinatorics on words, sequences with some required properties are usually looked up among morphic sequences. Let us recall what is known on letter frequencies and balancedness of such sequences. It clarifies why such sequences cannot be used in the proof of Theorem 2. Given two alphabets $\mathcal{A}$, $\mathcal{B}$, then a *morphism* is a map $\psi : \mathcal{A}^* \to \mathcal{B}^*$ such that $\psi(uv) = \psi(u)\psi(v)$ for all words $u, v \in \mathcal{A}^*$, where uv means concatenation of the words u and v. The morphism ψ can be naturally extended to a sequence $\mathbf{u} = u_0u_1u_2\cdots$ over $\mathcal{A}$ by setting $\psi(\mathbf{u}) = \psi(u_0)\psi(u_1)\psi(u_2)\cdots$.

A *fixed point* of a morphism $\psi : \mathcal{A}^* \to \mathcal{A}^*$ is a sequence $\mathbf{u}$ such that $\psi(\mathbf{u}) = \mathbf{u}$. We associate to a morphism $\psi : \mathcal{A}^* \to \mathcal{A}^*$ the *incidence matrix* M_ψ defined for each $i, j \in \{1, 2, \ldots, d\}$ as $(M_\psi)_{ij} = |\psi(j)|_i$. A morphism ψ is *primitive* if the matrix M_ψ is primitive, i.e., there exists $k \in \mathbb{N}$ such that M_ψ^k is a positive matrix.

In the sequel, we limit our consideration to the fixed points of primitive morphisms. Frequencies of letters in any fixed point of a primitive morphism are given by the coordinates of a unique positive eigenvector of norm one corresponding to the spectral radius [13]. Therefore, the letter frequencies belong to an algebraic field of order no greater than d. The same is true for morphic sequences, i.e., morphic images of fixed points. The letters in any morphic sequence $\mathbf{u}$ have the so-called *uniform letter frequencies* [13]: for any sequence $(k_n)_{n\in\mathbb{N}}$ of non-negative integers, the limit:

$$\lim_{n\to\infty} \frac{\#\{k_n \le i < k_n + n : u_i = a\}}{n}$$

exists and is the same for any choice of $(k_n)_{n\in\mathbb{N}}$. The definition of letter frequency in the introduction is restricted to the study of the limit for the sequence $(k_n) = (0)$.

The relation of balancedness and uniform letter frequency was described by Berthé and Delecroix [14].

Theorem 3. *A sequence $\mathbf{u}$ over an alphabet $\mathcal{A}$ is k-balanced for some $k \in \mathbb{N}$ if and only if it has uniform letter frequencies and there exists a constant B such that for any factor w of $\mathbf{u}$, we have $\big||w|_a - f_a|w|\big| \le B$ for every letter $a \in \mathcal{A}$.*

In general, proving that a sequence $\mathbf{u}$ is k-balanced for some k may be a complicated problem. Determine the minimum such k is even more difficult. If $\mathbf{u}$ is a fixed point of a primitive morphism, then it is possible to decide about k-balancedness using a result by Adamczewski [15]. It says that if all eigenvalues but one of the incidence matrix lie in the interior of the unit ball centred at the origin, then $\mathbf{u}$ is k-balanced. Moreover, an algorithm computing the minimum value of k is provided ibidem.

Since the letter frequencies in morphic sequences are always algebraic numbers, they cannot cover all possible candidates for $\vec{f}$.

A very important generalisation of morphic sequences is represented by S-adic systems, see for example [14]. An S-adic system introduced by Cassaigne in [16] allows constructing ternary sequences with prescribed letter frequencies that are almost always k-balanced for some constant k, as proven in [17].

3. Colouring of sequences

In this section, we describe a construction that enables us to create sequences with prescribed letter frequencies.

Definition 4. Let $\mathbf{u} = u_0u_1u_2\cdots$ be a sequence over the alphabet $\{a, b\}$. Denote by $\mathcal{O}_n^{(a)}$ and $\mathcal{O}_n^{(b)}$ the n^{th} occurrence of the letters a and b in $\mathbf{u}$, respectively. Let $\mathbf{a} = a_0a_1a_2\cdots$ and $\mathbf{b} = b_0b_1b_2\cdots$ be two sequences over two disjoint alphabets $\mathcal{A}$ and $\mathcal{B}$, respectively. Colouring of $\mathbf{u}$ by $\mathbf{a}$ and $\mathbf{b}$ is a sequence $\mathbf{v} = v_0v_1v_2\cdots$ over $\mathcal{A}\cup\mathcal{B}$ such that for every $N \in \mathbb{N}$ the N^{th} entry of $\mathbf{v}$ is:

$$v_N = \begin{cases} a_n, & \text{if } N = \mathcal{O}_n^{(a)}, \\ b_n, & \text{if } N = \mathcal{O}_n^{(b)}. \end{cases}$$

We denote $\mathbf{v} = \mathrm{colour}(\mathbf{u}, \mathbf{a}, \mathbf{b})$.

Less formally: $\mathbf{v}$ is obtained from the sequence $\mathbf{u}$ over $\{a, b\}$ by replacing the letters a's in $\mathbf{u}$ step by step by entries of the sequence $a_0a_1a_2\cdots$ and analogously, the letters b's in $\mathbf{u}$ are replaced by entries of the sequence $b_0b_1b_2\cdots$.

For $\mathbf{v} = \mathrm{colour}(\mathbf{u}, \mathbf{a}, \mathbf{b})$ we use the notation $\pi(\mathbf{v}) = \mathbf{u}$ and $\pi(v) = u$ for any $v \in \mathcal{L}(\mathbf{v})$ and the corresponding $u \in \mathcal{L}(\mathbf{u})$. We say that $\mathbf{u}$ (resp. u) is a *projection* of $\mathbf{v}$ (resp. v). The map $\pi : \mathcal{L}(\mathbf{v}) \to \mathcal{L}(\mathbf{u})$ is clearly a morphism.

Example 5. Let $\mathbf{u}$ be a sequence over $\{a, b\}$ and $\mathbf{a} = (121314)^\omega$ and $\mathbf{b} = (56)^\omega$, then:

$$\mathbf{u} = aabaababaabaababaababaabaababaabaab\cdots$$

$$\mathbf{v} = 125136154162153614516215316451261 35\cdots$$

We have $\pi(54162153614) = baabaababaa$.

Remark 6. Frequencies of letters in $\mathbf{v}=\text{colour}(\mathbf{u},\mathbf{a},\mathbf{b})$ can be easily computed from frequencies of letters in $\mathbf{u}$, $\mathbf{a}$, and $\mathbf{b}$: if they exist. For example, the letter $j \in \mathcal{A}$ has the frequency $f_a\gamma$ in $\mathbf{v}$, where f_a is the frequency of the letter a in $\mathbf{u}$ and γ is the frequency of the letter j in $\mathbf{a}$.

Definition 7. A sequence $\mathbf{a}$ is a *constant gap sequence* if for each letter a occurring in $\mathbf{a}$ the distance between any consecutive occurrences of a in $\mathbf{a}$ is constant.

Example 5 shows constant gap sequences $\mathbf{a}$ and $\mathbf{b}$. The next result comes from [18].

Lemma 8. *Let $\mathbf{a}$ be a constant gap sequence over an alphabet $\mathcal{A}$ containing more than one letter. Then $\mathbf{a}$ contains two distinct letters having the same frequency.*

Theorem 9 ([7])**.** *A recurrent aperiodic sequence $\mathbf{v}$ is 1-balanced if and only if $\mathbf{v} = \text{colour}(\mathbf{u},\mathbf{a},\mathbf{b})$ for some Sturmian sequence $\mathbf{u}$ and constant gap sequences $\mathbf{a},\mathbf{b}$ over two disjoint alphabets.*

Example 5 shows a 1-balanced sequence $\mathbf{v}$.

Using Remark 6, we can describe the form of the frequency vector in 1-balanced ternary and quaternary sequences.

Observation 10. *Let $\mathbf{v} = \text{colour}(\mathbf{u},\mathbf{a},\mathbf{b})$ be a 1-balanced sequence over an alphabet $\{1,2,\ldots,d\}$, where α is the frequency of a in $\mathbf{u}$. Then the frequency vector $\vec{f}_{\mathbf{v}}$ takes on the following values.*

(1.) *For $d=3$, we have only one frequency vector $\vec{f}_{\mathbf{v}} = (\alpha\frac{1}{2}, \alpha\frac{1}{2}, 1-\alpha)$ (up to certain letter permutations) corresponding to $\mathbf{a}=(12)^\omega$ and $\mathbf{b}=(3)^\omega$.*

(2.) *For $d=4$, we have three possibilities (up to certain letter permutations):*

- *if $\mathbf{a}=(12)^\omega$, $\mathbf{b}=(34)^\omega$, then:*
$$\vec{f}_{\mathbf{v}} = (\alpha\tfrac{1}{2}, \alpha\tfrac{1}{2}, (1-\alpha)\tfrac{1}{2}, (1-\alpha)\tfrac{1}{2}),$$
- *if $\mathbf{a}=(123)^\omega, \mathbf{b}=(4)^\omega$, then:*
$$\vec{f}_{\mathbf{v}} = (\alpha\tfrac{1}{3}, \alpha\tfrac{1}{3}, \alpha\tfrac{1}{3}, 1-\alpha),$$
- *if $\mathbf{a}=(1213)^\omega, \mathbf{b}=(4)^\omega$, then:*
$$\vec{f}_{\mathbf{v}} = (\alpha\tfrac{1}{2}, \alpha\tfrac{1}{4}, \alpha\tfrac{1}{4}, 1-\alpha).$$

The following lemma implies that $BT(d) \geq 2$ for $d \geq 3$.

Lemma 11. *Let $\mathbf{v}$ be a d-ary 1-balanced sequence, where $\mathbf{v} = \text{colour}(\mathbf{u},\mathbf{a},\mathbf{b})$, $\mathbf{u}$ is a Sturmian sequence over $\{a,b\}$ and $\mathbf{a}$, $\mathbf{b}$ are constant gap sequences over disjoint alphabets $\mathcal{A}$ and $\mathcal{B}$. If $d \geq 3$, then $\mathbf{v}$ contains two distinct letters of the same frequency.*

Proof. Denote α the frequency of the letter a in $\mathbf{u}$. If $d \geq 3$, then either $\#\mathcal{A} \geq 2$ or $\#\mathcal{B} \geq 2$. Consequently, by Lemma 8, either $\mathbf{a}$ or $\mathbf{b}$ contains two distinct letters i,j such that they have the same frequency in $\mathbf{a}$, resp. $\mathbf{b}$, say γ. Then, by Remark 6, $f_i = f_j = \alpha\gamma$, resp. $f_i = f_j = (1-\alpha)\gamma$, are the frequencies of letters i and j in $\mathbf{v}$. ∎

As a consequence of Lemma 11, we can see that the frequency vectors of 1-balanced sequences do not take on all possible values.

Lemma 12. *Let $\mathbf{u}$ be an ℓ-balanced sequence over $\{a,b\}$, and $\mathbf{a} = a_0a_1a_2\cdots$ and $\mathbf{b} = b_0b_1b_2\cdots$ be two k-balanced sequences over two disjoint alphabets $\mathcal{A}$ and $\mathcal{B}$, respectively. Then $\mathbf{v} = \text{colour}(\mathbf{u},\mathbf{a},\mathbf{b})$ is $(k+\ell)$-balanced.*

Proof. Let u, v be factors of $\mathbf{v}$ of the same length. We want to prove that for each letter $c \in \mathcal{A}\cup\mathcal{B}$:

$$\big||u|_c - |v|_c\big| \leq k+\ell\,.$$

WLOG let $c \in \mathcal{A}$. Denote $u' = \pi(u)$ and $v' = \pi(v)$. Clearly, $|u'| = |v'|$. Thanks to ℓ-balancedness of $\mathbf{u}$, we have $\big||u'|_a - |v'|_a\big| \leq \ell$. Let $\pi_{\mathcal{A}} : (\mathcal{A}\cup\mathcal{B})^* \to \mathcal{A}^*$ be a morphism such that $\pi_{\mathcal{A}}(x) = x$ if $x \in \mathcal{A}$ and $\pi_{\mathcal{A}}(x) = \varepsilon$ if $x \in \mathcal{B}$. It holds that for each factor w of $\mathbf{v}$, the word $\pi_{\mathcal{A}}(w)$ is a factor of $\mathbf{a}$. By definition of $\pi_{\mathcal{A}}$, we have $|u|_c = |\pi_{\mathcal{A}}(u)|_c$, $|v|_c = |\pi_{\mathcal{A}}(v)|_c$ and by definition of colouring, $|\pi_{\mathcal{A}}(u)| = |u'|_a$, $|\pi_{\mathcal{A}}(v)| = |v'|_a$.

Since $|u'|_a$ and $|v'|_a$ differ at most by ℓ, the words $\pi_{\mathcal{A}}(u)$ and $\pi_{\mathcal{A}}(v)$ are factors of $\mathbf{a}$ whose lengths differ at most by ℓ.

WLOG assume $|\pi_{\mathcal{A}}(u)| = |\pi_{\mathcal{A}}(v)| + n$, where $0 \leq n \leq \ell$. Then $\pi_{\mathcal{A}}(u) = a_i\cdots a_{i+m+n}$ and $\pi_{\mathcal{A}}(v) = a_j\cdots a_{j+m}$ for some $i,j,m \in \mathbb{N}$. Then using k-balancedness of $\mathbf{a}$ we get:

$$\begin{aligned}&\big||\pi_{\mathcal{A}}(u)|_c - |\pi_{\mathcal{A}}(v)|_c\big| \\ &\quad\leq \big||a_i\cdots a_{i+m}|_c - |a_j\cdots a_{j+m}|_c\big| \\ &\qquad + |a_{i+m+1}\cdots a_{i+m+n}|_c \\ &\quad\leq k+n \leq k+\ell.\end{aligned}$$

Since $|u|_c = |\pi_{\mathcal{A}}(u)|_c$ and $|v|_c = |\pi_{\mathcal{A}}(v)|_c$, we have proven that $\big||u|_c - |v|_c\big| \leq k+\ell$. ∎

4. Proof of Theorem 2

In this section, we prove a statement having Theorem 2 as its direct consequence. We make use of the knowledge of the number of occurrences of letters in Sturmian sequences, provided in [19].

Lemma 13. *Let $\mathbf{u}$ be a 1-balanced sequence over the alphabet $\{a,b\}$ and $f_a = \alpha \in (0,1)$. Then any factor u of length $n \in \mathbb{N}$ either contains $\lceil\alpha n\rceil$ letters a and $\lfloor(1-\alpha)n\rfloor$ letters b, or u contains $\lfloor\alpha n\rfloor$ letters a and $\lceil(1-\alpha)n\rceil$ letters b.*

Theorem 14. *Let $d \in \mathbb{N}$, $d \geq 1$, and $f(1)$, $f(2),\ldots,f(d)$ be positive numbers such that $f(1)+f(2)+\cdots+f(d)=1$. Then there exists an infinite sequence $\mathbf{v}$ over the alphabet $\{1,2,\ldots,d\}$ such that:*

(1.) *the frequency of the letter i in $\mathbf{v}$ is $f(i)$ for each $i \in \{1,2,\ldots,d\}$,*

(2.) *$\mathbf{v}$ is k-balanced with $k = \lceil\log_2 d\rceil$,*

(3.) *the factor complexity of* $\mathbf{v}$ *satisfies:*

$$\mathcal{C}_{\mathbf{v}}(n) \leq (n+1)^{d-1}.$$

Proof. We proceed by induction on d. If $d = 1$, then we put $\mathbf{v} = 1^\omega$.

Let $d \geq 2$. We denote $\alpha = f(1)+f(2)+\cdots+f(\lceil \frac{d}{2} \rceil)$ and:

$$f'(i) = \begin{cases} \dfrac{1}{\alpha} f(i) & \text{for } i = 1, 2, \ldots, \lceil \dfrac{d}{2} \rceil, \\ \dfrac{1}{1-\alpha} f(i) & \text{for } i = \lceil \dfrac{d}{2} \rceil + 1, \ldots, d. \end{cases}$$

Obviously:

$$f'(1)+\cdots+f'(\lceil \tfrac{d}{2} \rceil) \;=\; 1 \;=\; f'(\lceil \tfrac{d}{2} \rceil+1)+\cdots+f'(d).$$

By induction hypothesis, there exist a k_a-balanced sequence $\mathbf{a} = a_0a_1a_2\cdots$ over the alphabet $\mathcal{A} = \{1, 2, \ldots, \lceil \frac{d}{2} \rceil\}$ with the frequencies of letters $f'(i)$ for each $i \in \mathcal{A}$ and $k_a = \lceil \log_2 \lceil \frac{d}{2} \rceil \rceil$ and a k_b-balanced sequence $\mathbf{b} = b_0b_1b_2\cdots$ over the alphabet $\mathcal{B} = \{\lceil \frac{d}{2} \rceil + 1, \ldots, d\}$ with the frequencies of letters $f'(i)$ for each $i \in \mathcal{B}$ and $k_b = \lceil \log_2 \lfloor \frac{d}{2} \rfloor \rceil$.

Let $\mathbf{u}$ be a 1-balanced sequence over the alphabet $\{a, b\}$ with frequencies of letters α and $1-\alpha$, respectively. Then the sequence $\mathbf{v} = \mathrm{colour}(\mathbf{u}, \mathbf{a}, \mathbf{b})$ is over the alphabet $\{1, \ldots, d\}$, and by Remark 6, the frequencies of letters are $\alpha f'(i) = f(i)$ for $i \in \mathcal{A}$ and $(1-\alpha)f'(i) = f(i)$ for $i \in \mathcal{B}$.

Lemma 12 implies that $\mathbf{v} = \mathrm{colour}(\mathbf{u}, \mathbf{a}, \mathbf{b})$ is k-balanced, with $k = 1 + \lceil \log_2 \lceil \frac{d}{2} \rceil \rceil$. To complete the proof of Item (2.), we have to show that $1 + \lceil \log_2 \lceil \frac{d}{2} \rceil \rceil \leq \lceil \log_2 d \rceil$. If d is even, then $1 + \lceil \log_2 \lceil \frac{d}{2} \rceil \rceil = \lceil \log_2 d \rceil$.

For d odd, we demonstrate the required inequality by contradiction. Assume that $1 + \lceil \log_2 \lceil \frac{d}{2} \rceil \rceil > \lceil \log_2 d \rceil$. As d is odd, we know that:

$$\begin{aligned} \log_2 d &< \lceil \log_2 d \rceil \leq \lceil \log_2 \lceil \tfrac{d}{2} \rceil \rceil \\ &= \lceil \log_2 \tfrac{d+1}{2} \rceil = \lceil \log_2 (d+1) \rceil - 1 < \log_2 (d+1). \end{aligned}$$

Therefore we have $d < 2^{\lceil \log_2 d \rceil} < d + 1$. The numbers $d, 2^{\lceil \log_2 d \rceil}$ and $d + 1$ are integers, which leads to a contradiction.

To show Item (3.), we proceed again by induction on d. Let u be a fixed factor of length n in the 1-balanced sequence $\mathbf{u}$. The frequencies of letters in $\mathbf{u}$ are α and $(1-\alpha)$. By Lemma 13, the factor u either contains $\lceil \alpha n \rceil$ letters a and $\lfloor (1-\alpha)n \rfloor$ letters b, or u contains $\lfloor \alpha n \rfloor$ letters a and $\lceil (1-\alpha)n \rceil$ letters b. Hence the factor u equals the projection $\pi(v)$ for at most $\mathcal{C}_{\mathbf{a}}(\lceil \alpha n \rceil) \times \mathcal{C}_{\mathbf{b}}(\lceil (1-\alpha)n \rceil)$ factors v in $\mathbf{v}$. Since $\mathbf{u}$ has at most $n+1$ factors of length n, we have:

$$\mathcal{C}_{\mathbf{v}}(n) \leq (n+1)\, \mathcal{C}_{\mathbf{a}}(\lceil \alpha n \rceil)\, \mathcal{C}_{\mathbf{b}}(\lceil (1-\alpha)n \rceil).$$

Using the induction hypothesis and simple inequalities $\lceil \alpha n \rceil \leq n$ and $\lceil (1-\alpha)n \rceil \leq n$, we conclude:

$$\begin{aligned} \mathcal{C}_{\mathbf{v}}(n) &\leq (n+1)\, \mathcal{C}_{\mathbf{a}}(n)\, \mathcal{C}_{\mathbf{b}}(n) \\ &\leq (n+1)\, (n+1)^{\lceil \frac{d}{2} \rceil - 1}\, (n+1)^{\lfloor \frac{d}{2} \rfloor - 1} \\ &= (n+1)^{d-1}. \end{aligned}$$

∎

5. Comments and questions

(1.) The bound we found on the factor complexity of the sequence $\mathbf{v}$ constructed in the proof of Theorem 14 is not optimal. What is the optimal upper bound?

(2.) A 1-balanced binary sequence $\mathbf{u}$ is either Sturmian or periodic. If some ratio $f(i) : f(j)$ in the assumptions of Theorem 14 is rational, we can reduce the degree of the polynomial in the upper bound $(n+1)^{d-1}$ at least by 1.

(3.) When all frequencies $f(i)$ in the assumptions of Theorem 14 are rational, our construction gives a periodic sequence $\mathbf{v}$. How to determine its period?

(4.) Given rational frequencies $f(i) = \frac{p_i}{q_i}$, how many steps are needed to construct a prefix of $\mathbf{v}$ of length N?

(5.) A *cubic billiard sequence in dimension* d (also called hypercubic billiard sequence) is a coding of the sequence of the faces successively hit by a billiard ball moving inside the unit hypercube $[0, 1]^d$, where two parallel faces are encoded by the same letter. These d-ary sequences are parametrised by the initial position $x \in [0, 1]^d$ and the initial momentum $\theta \in R^d \setminus \{0\}$ of the ball. The vector of letter frequencies corresponds to the vector of initial momentum, up to a dilatation, and a change in the signs of certain components. Under an additional condition on momentum, the factor complexity of cubic billiard sequences in dimension d satisfies $\mathcal{C}(n) = n^{d-1}(1 + o(1))$, see [20]. Vuillon [21] proved that any cubic billiard sequence in dimension d, whose momentum has rationally independent components, is $d-1$ balanced. Andrieu and Vivion [9] further specified that:

- for $d \in \{2, 3, 4\}$, any cubic billiard sequence in dimension d generated by a momentum with rationally independent components is not $d-2$ balanced,
- for $d \geq 5$, for every $C \in \{3, 4, \ldots, d-1\}$, there exists a cubic billiard sequence in dimension d generated by a momentum with rationally independent components that is C-balanced, but not $(C-1)$-balanced.

We can conclude that for certain frequency vectors $\vec{f} \in \mathbb{R}^d$, the corresponding cubic billiard sequence in dimension d is $(d-1)$-balanced but not $(d-2)$-balanced, while for the same $\vec{f}$, the colouring procedure we use in the proof of Theorem 14

provides a sequence which is $\lceil \log_2 d \rceil$-balanced. In other words, in the case of $d > 3$, we construct a sequence which is less imbalanced.

Acknowledgements

Both authors are aware of how significantly Prof. Havlíček influenced their scientific careers. As a teacher, as a co-author, and most importantly, as a person who, in his role as dean of the faculty and head of the Department of Mathematics, encouraged his colleagues to pursue scientific endeavours.

References

[1] M. Morse, G. A. Hedlund. Symbolic dynamics II. Sturmian trajectories. *American Journal of Mathematics* **62**(1):1–42, 1940. https://doi.org/10.2307/2371431

[2] L'. Balková, E. Pelantová, Š. Starosta. Sturmian jungle (or garden?) on multiliteral alphabets. *RAIRO – Theoretical Informatics and Applications* **44**(4):443–470, 2010. https://doi.org/10.1051/ita/2011002

[3] G. Richomme, K. Saari, L. Q. Zamboni. Abelian complexity of minimal subshifts. *Journal of the London Mathematical Society* **83**(1):79–95, 2011. https://doi.org/10.1112/jlms/jdq063

[4] L. Vuillon. A characterization of Sturmian words by return words. *European Journal of Combinatorics* **22**(2):263–275, 2001. https://doi.org/10.1006/eujc.2000.0444

[5] P. Arnoux, Š. Starosta. The Rauzy gasket. In B. Boston (ed.), *Further Developments in Fractals and Related Fields*, Trends in Mathematics, pp. 1–23. Birkhäuser, Boston, 2013. https://doi.org/10.1007/978-0-8176-8400-6_1

[6] A. Zorich. Deviation for interval exchange transformations. *Ergodic Theory and Dynamical Systems* **17**(6):1477–1499, 1997. https://doi.org/10.1017/S0143385797086215

[7] P. Hubert. Suites équilibrées. *Theoretical Computer Science* **242**(1–2):91–108, 2000. https://doi.org/10.1016/S0304-3975(98)00202-3

[8] L. Dvořáková, Z. Masáková, E. Pelantová. 2-balanced sequences coding rectangle exchange transformation. *Theoretical Computer Science* **68**(6):1537–1555, 2024. https://doi.org/10.1007/s00224-024-10188-6

[9] M. Andrieu, L. Vivion. Imbalances in hypercubic billiard words. In *18th Mons Theoretical Computer Science Days*. Prague, 2022.

[10] P. Arnoux, C. Mauduit, I. Shiokawa, J.-I. Tamura. Complexity of sequences defined by billiard in the cube. *Bulletin de la Société Mathématique de France* **122**(1):1–12, 1994. https://doi.org/10.24033/bsmf.2220

[11] P. Arnoux, C. Mauduit, I. Shiokawa, J.-I. Tamura. Rauzy's conjecture on billiards in the cube. *Tokyo Journal of Mathematics* **17**(1):211–218, 1994. https://doi.org/10.3836/tjm/1270128200

[12] N. Bedaride, P. Hubert. Billiard complexity in the hypercube. *Annales de l'Institut Fourier* **57**(3):719–738, 2007. https://doi.org/10.5802/aif.2274

[13] M. Queffélec. *Substitution dynamical systems – spectral analysis*. Lecture notes in mathematics. Springer-Verlag, Heidelberg, 2nd edn., 2010. https://doi.org/10.1007/978-3-642-11212-6

[14] V. Berthé, V. Delecroix. Beyond substitutive dynamical systems: S-adic expansions. *RIMS Kôkyûroku Bessatsu* **B46**:81–123, 2014.

[15] B. Adamczewski. Balances for fixed points of primitive substitutions. *Theoretical Computer Science* **307**(1):47–75, 2003. https://doi.org/10.1016/S0304-3975(03)00092-6

[16] J. Cassaigne. Un algorithme de fractions continues de complexité linéaire, 2015. DynA3S meeting, LIAFA, Paris.

[17] J. Cassaigne, S. Labbé, J. Leroy. Almost everywhere balanced sequences of complexity $2n + 1$. *Combinatorics and Number Theory* **11**(4):287–333, 2022. https://doi.org/10.2140/moscow.2022.11.287

[18] M. Newman. Roots of unity and covering sets. *Mathematische Annalen* **191**(4):279–282, 1971. https://doi.org/10.1007/BF01350330

[19] M. Lothaire. *Algebraic combinatorics on words*. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, Cambridge, 2002. https://doi.org/10.1017/CBO9781107326019

[20] N. Bedaride. Directional complexity of the hypercubic billiard. *Discrete Mathematics* **309**(8):2053–2066, 2009. https://doi.org/10.1016/j.disc.2008.04.018

[21] L. Vuillon. Balanced words. *Bulletin of the Belgian Mathematical Society – Simon Stevin* **10**(5):787–805, 2003. https://doi.org/10.36045/bbms/1074791332