Factor-balanced S-adic languages

Léo Poirier, Wolfgang Steiner

FACTOR-BALANCED S-ADIC LANGUAGES

LÉO POIRIER AND WOLFGANG STEINER

Abstract. A set of words, also called a language, is letter-balanced if the number of occurrences of each letter only depends on the length of the word, up to a constant. Similarly, a language is factor-balanced if the difference of the number of occurrences of any given factor in words of the same length is bounded. The most prominent example of a letter-balanced but not factor-balanced language is given by the Thue–Morse sequence. We establish connections between the two notions, in particular for languages given by substitutions and, more generally, by sequences of substitutions. We show that the two notions essentially coincide when the sequence of substitutions is proper. For the class of Thue–Morse–Sturmian languages, we give a full characterisation of factor-balancedness.

1. Introduction

The study of balancedness of languages goes back at least to Morse and Hedlund [MH40] who proved that each block of length n in a Sturmian sequence of slope α has $\lfloor n\alpha \rfloor$ or $\lceil n\alpha \rceil$ occurrences of the letter that has frequency α (and thus $\lfloor n(1-\alpha) \rfloor$ or $\lceil n(1-\alpha) \rceil$ occurrences of the other letter). In other words, the difference between the number of occurrences of a letter in blocks of the same length is at most 1; we call this property letter-1-balanced, previously it has often been simply called balanced. More generally, a language is letter-balanced if the number of occurrences of a letter only depends on the length of a word in the language, up to an additive constant. We do not only consider the occurrence of letters but also of longer blocks, and we say that a language is factor-balanced if the number of occurrences of each block in a word of the language only depends on the length of the word, up to a constant that can depend on the block. Usually, languages coming from a symbolic dynamical system (or subshift) are considered; we use the slightly weaker property of being factorial. For infinite sequences, balancedness is equivalent to bounded symbolic discrepancy, as studied in [Ada04]. These concepts have applications in operations research, for optimal routing and scheduling and are related to Fraenkel's conjecture; see [BCB19] for references.

Some relations between letter-balancedness and factor-balancedness have been studied in [Ada03, Que10], and more recently in [BCB19]. Here, we improve on these results and show that factor-balancedness is preserved by the application of a substitution. We consider languages (or subshifts) given by a sequence of substitutions, also called S-adic languages. When the substitutions are left or right proper, i.e., the image of each letter starts or ends with the same letter, we show that letter-balancedness (on all levels) implies factor-balancedness; this was previously known only under the assumption that the substitutions have unimodular incidence matrices [BCBD$^+$21]. A particular case is that of a substitutive shift with a proper substitution. Here, we cannot remove the assumption of properness, as for example the

Date: November 23, 2023.

This work was supported by the Agence Nationale de la Recherche through the projects CODYS (ANR-18-CE40-0007) and IZES (ANR-22-CE40-0011).

Thue–Morse shift is not factor-balanced [Sad16, BCB19]; we give a short proof in Section 5. For non-proper substitutions, we have to require balancedness for all factors of length 2, not only for letters, in order to get factor-balancedness.

In Section 2, we define most of the notions and give a characterisation of letter-balancedness in terms of the distance to a frequency vector. The effect of substitutions on balancedness is studied in Section 3. Section 4 contains our main results, on sequences of substitutions. Finally, we consider the balancedness of a particular class of S-adic languages in Section 5.

2. Balancedness

For a finite alphabet A, let A^* be the set of finite words over A. A *language* is a subset $\mathcal{L} \subseteq A^*$. A word $v \in A^*$ is a *factor* of $w \in A^*$ if there exist $p, s \in A^*$ such that $w = pvs$; here, p is a *prefix* and s is a *suffix* of w. We denote the set of factors of w by $\mathcal{F}(w)$, and the set of factors of elements of $\mathcal{L}$ by $\mathcal{F}(\mathcal{L})$. A langage $\mathcal{L}$ is *factorial* if $\mathcal{F}(\mathcal{L}) = \mathcal{L}$. The length of a word $w \in A^*$ is denoted by $|w|$, i.e., $|w| = n$ if $w \in A^n$. We denote the prefix (resp. suffix) of length n of a word w by $\mathrm{pref}_n(w)$ (resp. $\mathrm{suff}_n(w)$). The number of *occurrences* of a word v in w, i.e., the number of different decompositions $w = pvs$, is denoted by $|w|_v$. A language $\mathcal{L}$ is called *C-balanced w.r.t. v* if

$$|w|_v - |w'|_v \le C \qquad \text{for all } w, w' \in \mathcal{L} \text{ with } |w| = |w'|.$$

(By symmetry, this is equivalent to $\big||w|_v - |w'|_v\big| \le C$ for all $w, w' \in \mathcal{L}$ with $|w| = |w'|$.) It is called *C-balanced for length n* if it is C-balanced for all $v \in A^n$. We often omit the C and say that a language is *balanced for length n* if it is C-balanced for length n for some $C \ge 0$. Instead of "(C-)balanced for length 1", we also say *letter-(C-)balanced*; other papers use the term "balanced" instead of letter-1-balanced or instead of letter-balanced, and a letter-balanced language is sometimes called "finitely balanced". A language is *factor-(C-)balanced* if it is (C-)balanced for all lengths $n \ge 1$; in [Sad16], a factor-balanced language is called *totally finitely balanced*. Note that factor-C-balancedness is a strong property that is satisfied for certain Sturmian languages [FV02], but we do not study this property here. We are only interested in factor-balancedness, which means that for each n there exists C_n such that $\mathcal{L}$ is C_n-balanced for length n; equivalently, for each $v \in \mathcal{F}(\mathcal{L})$ there exists C_v such that $\mathcal{L}$ is C_v-balanced w.r.t. v. (Note that $|w|_v = 0$ for all $w \in \mathcal{L}$, $v \notin \mathcal{F}(\mathcal{L})$.) We first show that balancedness for length n is the same as balancedness for lengths up to n.

Lemma 2.1. *If a language is balanced for length n, then it is balanced for all lengths $k \le n$.*

Proof. Let $\mathcal{L} \subset A^*$ be C-balanced for length n, $k < n$. For all $v \in A^k$, $w \in A^*$, we have $|w|_v = \sum_{s \in A^{n-k}} |w|_{vs} + |\mathrm{suff}_{n-1}(w)|_v$, thus $|w|_v - |w'|_v \le (\#A)^{n-k} C + n - 1$ for $w' \in A^{|w|}$. □

We will also use the following characterisation of letter-balancedness in terms of distance from the line defined by a frequency vector, cf. [BT02, Ada03].

Proposition 2.2. *Let $\mathcal{L} \subset A^*$ be an infinite letter-C-balanced factorial language. Then there exists a (frequency) vector $(f_a)_{a \in A}$ such that $\big||w|_a - f_a|w|\big| \le C$ for all $a \in A$, $w \in \mathcal{L}$.*

Proof. Since $[0,1]^{\#A}$ is compact, there exists a vector $(f_a)_{a \in A}$ and a sequence of words $v_n \in \mathcal{L}$ such that $\lim_{n\to\infty} |v_n| = \infty$ and $\lim_{n\to\infty} |v_n|_a / |v_n| = f_a$ for all $a \in A$. For arbitrary but fixed

$w \in \mathcal{L}$, set $k_n = \lfloor |v_n|/|w| \rfloor$, and decompose $v_n = v_{n,1} \cdots v_{n,k_n} v_{n,k_n+1}$ with $|v_{n,i}| = |w|$ for all $1 \le i \le k_n$. Since $v_{n,i} \in \mathcal{L}$, $\mathcal{L}$ is letter-C-balanced, and $|v_{n,k_n+1}| < |w|$, we obtain that

$$\left| |w|_a - \frac{|v_n|_a}{|v_n|} |w| \right| = \frac{|w|}{|v_n|} \left| \frac{|v_n|}{|w|} |w|_a - |v_n|_a \right| < \frac{|w|}{|v_n|} \left(|w| + \sum_{i=1}^{k_n} \big| |w|_a - |v_{n,i}|_a \big| \right) \le \frac{|w|^2}{|v_n|} + \frac{k_n |w|}{|v_n|} C$$

for all $a \in A$. Letting $n \to \infty$, this gives that $\big| |w|_a - f_a |w| \big| \le C$. □

Lemma 2.3. *Let $\mathcal{L} \subset A^*$ and $(f_a)_{a \in A}$ such that $\big| |w|_a - f_a |w| \big| \le C$ for all $a \in A$, $w \in \mathcal{L}$. Then $\mathcal{L}$ is letter-$(2C)$-balanced.*

Proof. We have $|w|_a - |w'|_a = |w|_a - f_a|w| + f_a|w'| - |w'|_a \le 2C$ for all $w, w' \in \mathcal{L}$, $a \in A$, such that $|w| = |w'|$. □

3. Substitutions

In this section, we study how the application of a substitution influences balancedness. Here, a *substitution* σ is a morphism from A^* to B^*, with the operation of concatenation, i.e., $\sigma(vw) = \sigma(v)\sigma(w)$ for all $v, w \in A^*$. We use the notation

$$\|\sigma\| := \max_{a \in A} |\sigma(a)|, \qquad \langle \sigma \rangle := \min_{a \in A} |\sigma(a)|,$$

and call a substitution *non-erasing* if all images of letters are non-empty, i.e., $\langle \sigma \rangle \ge 1$. It is left (resp. right) *proper* when all letter images start (resp. end) with the same letter.

To show that substitutions preserve balancedness, we use the following lemma.

Lemma 3.1. *Let $\mathcal{L} \subset A^*$ be a letter-C-balanced factorial language and $\sigma : A^* \to B^*$ a substitution. Then, for all $w, w' \in \mathcal{F}(\sigma(\mathcal{L}))$ with $|w| = |w'|$, there exist $x, x', z, z' \in B^*$, $y, y' \in \mathcal{L}$, such that*

$$w = x\,\sigma(y)\,z,\ w' = x'\,\sigma(y')\,z', \quad |y| = |y'|,\ \max\{|x\,z|, |x'z'|\} \le (2 + C\,\#A)\,\|\sigma\| - 2.$$

Proof. Since $w, w' \in \mathcal{F}(\sigma(\mathcal{L}))$ and $\mathcal{L}$ is factorial, we can write $w = x\,\sigma(v)\,u$ and $w = x'\sigma(v')\,u'$ with $v, v' \in \mathcal{L}$, $u, u', x, x' \in A^*$ such that $|u|, |u'|, |x|, |x'| < \|\sigma\|$. Assume w.l.o.g. that $|v| \le |v'|$, let $y = v$, $v' = y's'$ with $|y'| = |y|$, $z = u$, $z' = \sigma(s')\,u'$. Since $\mathcal{L}$ is factorial, we have $y, y' \in \mathcal{L}$. Since $\mathcal{L}$ is letter-C-balanced, we have

$$|\sigma(y)| - |\sigma(y')| = \sum_{a \in A} \big(|y|_a - |y'|_a \big)\, |\sigma(a)| \le (\#A)\, C\, \|\sigma\|,$$

thus

$$|x'z'| = |w'| - |\sigma(y')| \le |w| - |\sigma(y)| + (\#A)\, C\, \|\sigma\| \le 2\,(\|\sigma\| - 1) + (\#A)\, C\, \|\sigma\|.$$

Therefore, $w = x\,\sigma(y)\,z$ and $w' = x'\,\sigma(y')\,z'$ satisfy all the required properties. □

Proposition 3.2. *Let $\mathcal{L} \subset A^*$ be a factorial language and $\sigma : A^* \to B^*$ a substitution. If $\mathcal{L}$ is letter-balanced, then $\mathcal{F}(\sigma(\mathcal{L}))$ is letter-balanced.*

Proof. Suppose that $\mathcal{L}$ is C-letter-balanced, and let $w = x\,\sigma(y)\,z$, $w' = x'\,\sigma(y')\,z'$ be as in Lemma 3.1. Then, for all $b \in B$,

$$\begin{aligned} |w|_b - |w'|_b &= |xz|_b - |x'z'|_b + |\sigma(y)|_b - |\sigma(y')|_b \\ &\le (2 + C\,\#A)\,\|\sigma\| - 2 + \sum_{a \in A} \big(|y|_a - |y'|_a \big)\, |\sigma(a)|_b \le 2\,(1 + C\,\#A)\,\|\sigma\| - 2. \end{aligned}$$ □

To study balancedness for length n, we use the *n-th higher block code* of a word $a_1a_2\cdots a_N \in A^N$, $N \geq 0$, which is the word over the alphabet A^n defined by

$$(a_1a_2\cdots a_N)^{(n)} = (a_1a_2\cdots a_n)(a_2a_3\cdots a_{n+1})\cdots(a_{N-n+1}a_{N-n+2}\cdots a_N) \in (A^n)^{N-n+1}$$

if $N \geq n$, the empty word if $N < n$. Note that $|w|_v = |w^{(n)}|_v$ for all $v \in A^n$; in particular, a language $\mathcal{L}$ is C-balanced for length n if and only if $\{w^{(n)} : w \in \mathcal{L}\}$ is letter-C-balanced (over the alphabet A^n).

Proposition 3.3. *Let $\mathcal{L} \subset A^*$ be a factorial language that is balanced for length n, $\sigma : A^* \to B^*$ a substitution, and $u \in B^*$ a (possibly empty) word that is a prefix of $\sigma(a)u$ for all $a \in A$ or a suffix of $u\sigma(a)$ for all $a \in A$. Then $\mathcal{F}(\sigma(\mathcal{L}))$ is balanced for length $\min_{w\in\mathcal{L}\cap A^{n-1}} |\sigma(w)|+|u|+1$. In particular,*

- *$\mathcal{F}(\sigma(\mathcal{L}))$ is balanced for length n if σ is non-erasing,*
- *$\mathcal{F}(\sigma(\mathcal{L}))$ is balanced for length $n+1$ if σ is left or right proper.*

Proof. Let $\mathcal{L} \subset A^*$, $\sigma : A^* \to B^*$, $u \in B^*$ be as in the statement of the proposition, $1 \leq m \leq \min_{w\in\mathcal{L}\cap A^{n-1}} |\sigma(w)|+|u|+1$. Assume w.l.o.g. that u is a prefix of $\sigma(a)u$ for all $a \in A$, the suffix case being symmetric. We define a substitution $\hat{\sigma} : (A^n \cap \mathcal{L})^* \to (B^m)^*$ by

$$\hat{\sigma}(a_1a_2\cdots a_n) := \big(\sigma(a_1)\mathrm{pref}_{m-1}(\sigma(a_2\cdots a_n)u)\big)^{(m)} \quad \text{for } a_1\cdots a_n \in A^n \cap \mathcal{L}.$$

(Here, $\hat{\sigma}$ is a substitution on the alphabet $A^n \cap \mathcal{L}$, and the condition on m ensures that $\mathrm{pref}_{m-1}(\sigma(a_2\cdots a_n)u)$ exists.) Then we have, for all $w \in \mathcal{L}$,

$$\big(\sigma(w)u\big)^{(m)} = \hat{\sigma}\big(w^{(n)}\big)\,\big(\sigma(\mathrm{suff}_{n-1}(w))u\big)^{(m)}. \tag{3.1}$$

Let C_n be such that $\mathcal{L}$ is C_n-balanced for length n. Then, by Lemma 2.1, $\mathcal{L}$ is letter-C_1-balanced for some $C_1 \geq 0$. Let $w, w' \in \mathcal{F}(\sigma(\mathcal{L}))$ with $|w| = |w'|$, and write $w = x\,\sigma(y)\,z$, $w' = x'\,\sigma(y')\,z'$ as in Lemma 3.1. Similarly to (3.1), we obtain that

$$(wu)^{(m)} = \big(x\,\mathrm{pref}_{m-1}(\sigma(y)u)\big)^{(m)}\,\hat{\sigma}\big(y^{(n)}\big)\,\big(\sigma(\mathrm{suff}_{n-1}(y))zu\big)^{(m)}.$$

Using a similar decomposition for $(w'u)^{(m)}$, we obtain for $v \in B^m$ that

$$\begin{aligned}|w|_v - |w'|_v &\leq \max\big\{|xz|, |x'z'|\big\} + (n-1)\|\sigma\| + \sum_{t\in A^n\cap\mathcal{L}} \big(|y|_t - |y'|_t\big)\,\big|\hat{\sigma}(t)\big|_v \\ &\leq (2 + C_1\#A)\|\sigma\| - 2 + (n-1)\|\sigma\| + (\#A)^n\,C_n\,\|\sigma\|.\end{aligned}$$

Here, we have used that $|w|_v = |w^{(m)}|_v$, that $\mathcal{L}$ is C_n-balanced for length n, and that $|\hat{\sigma}(a_1\cdots a_n)| = |\sigma(a_1)|$ for $a_1\cdots a_n \in A^n \cap \mathcal{L}$. This proves that $\mathcal{F}(\sigma(\mathcal{L}))$ is balanced for length $\min_{w\in\mathcal{L}\cap A^{n-1}} |\sigma(w)|+|u|+1$. If σ is non-erasing, then $|\sigma(w)| \geq n-1$ for all $w \in A^{n-1}$, thus $\mathcal{F}(\sigma(\mathcal{L}))$ is balanced for length n. If σ is left or right proper, then σ is non-erasing and $|u| \geq 1$, thus $\mathcal{F}(\sigma(\mathcal{L}))$ is balanced for length $n+1$. □

Theorem 3.4. *Let $\mathcal{L} \subset A^*$ be a factorial language and $\sigma : A^* \to B^*$ a substitution. If $\mathcal{L}$ is factor-balanced, then $\mathcal{F}(\sigma(\mathcal{L}))$ is factor-balanced.*

Proof. For non-erasing σ, the theorem is a direct consequence of Proposition 3.3. If $\mathcal{F}(\sigma(\mathcal{L}))$ is finite, then it is also factor-balanced. If $\mathcal{F}(\sigma(\mathcal{L}))$ is infinite, then there exists $a \in A$ such that $|\sigma(a)| \geq 1$ and $\{|w|_a : w \in \mathcal{L}\}$ is unbounded. If $\mathcal{L}$ is letter-C-balanced, then Proposition 2.2 gives some $f_a \geq 0$ such that $|w|_a \geq f_a|w| - C$ and thus $|\sigma(w)| \geq f_a|w| - C$ for all $w \in \mathcal{L}$. By Proposition 3.3, balancedness of $\mathcal{L}$ for length n implies balancedness of

$\mathcal{F}(\sigma(\mathcal{L}))$ for length $f_a\,(n-1)-C+1$. Note that $f_a > 0$ since $|w|_a \le f_a|w| + C$ and $|w|_a$ is unbounded. Therefore, factor-balancedness of $\mathcal{L}$ implies that of $\mathcal{F}(\sigma(\mathcal{L}))$. □

When the incidence matrix of σ is invertible, we can also infer letter-balancedness of $\mathcal{L}$ from that of $\mathcal{F}(\sigma(\mathcal{L}))$. Here, the *incidence matrix* of a substitution $\sigma : A^* \to B^*$ is

$$\mathcal{M}_\sigma := (|\sigma(a)|_b)_{b\in B, a\in A}.$$

Proposition 3.5. *Let $\sigma : A^* \to B^*$ be a substitution with invertible incidence matrix $\mathcal{M}_\sigma$ and $\mathcal{L} \subset A^*$. If $\mathcal{F}(\sigma(\mathcal{L}))$ is letter-balanced, then $\mathcal{L}$ is letter-balanced.*

Proof. If $\mathcal{L}$ is finite, then it is trivially letter-balanced. Assume in the following that $\mathcal{L}$ is infinite and $\mathcal{M}_\sigma$ invertible. Then σ is non-erasing and thus $\sigma(\mathcal{L})$ infinite. If $\mathcal{F}(\sigma(\mathcal{L}))$ is letter-balanced, then Proposition 2.2 gives a frequency vector $\mathbf{f} = (f_b)_{b\in B}$ such that

$$D := \big\{(|\sigma(w)|_b - |\sigma(w)|f_b)_{b\in B} \,:\, w \in \mathcal{L}\big\}$$

is a bounded set. Since $(|\sigma(w)|_b)_{b\in B} = \mathcal{M}_\sigma(|w|_a)_{a\in A}$ and $\mathcal{M}_\sigma$ is invertible, we have

$$\mathcal{M}_\sigma^{-1}D = \big\{(|w|_a)_{a\in A} - |\sigma(w)|\mathcal{M}_\sigma^{-1}\mathbf{f} \,:\, w \in \mathcal{L}\}.$$

Therefore, the vectors $(|w|_a)_{a\in A}$, $w \in \mathcal{L}$, have bounded distance from the line $\mathbb{R}\mathcal{M}_\sigma^{-1}\mathbf{f}$. Hence, $\mathcal{M}_\sigma^{-1}\mathbf{f}$ is a non-negative vector and $(f'_a)_{a\in A} := \mathcal{M}_\sigma^{-1}\mathbf{f}/\|\mathcal{M}_\sigma^{-1}\mathbf{f}\|_1$ is the frequency vector of $\mathcal{L}$. Since $\sum_{a\in A}\big(|w|_a - |w|f'_a\big) = |w| - |w| = 0$ for all $w \in A^*$, the set

$$D' := \big\{(|w|_a - |w|f'_a)_{a\in A} \,:\, w \in \mathcal{L}\big\} = \Big\{(|w|_a)_{a\in A} - \frac{|w|}{\|\mathcal{M}_\sigma^{-1}\mathbf{f}\|_1}\mathcal{M}_\sigma^{-1}\mathbf{f} \,:\, w \in \mathcal{L}\Big\}$$

lies in the hyperplane $H := \{(x_a)_{a\in A} : \sum_{a\in A} x_a = 0\}$. Therefore, D' is the projection of $\mathcal{M}_\sigma^{-1}D$ to H along the line $\mathbb{R}\mathcal{M}_\sigma^{-1}\mathbf{f}$ (which is not in H since $\mathcal{M}_\sigma^{-1}\mathbf{f}$ is non-negative and non-zero), thus D' is bounded. Hence, by Lemma 2.3, $\mathcal{L}$ is letter-balanced. □

4. S-adic languages

Now, we consider sequences of substitutions $\boldsymbol{\sigma} = (\sigma_k)_{k\ge 0}$, $\sigma_k : A_{k+1}^* \to A_k^*$. We set

$$\sigma_{[k,n)} := \sigma_k \circ \sigma_{k+1} \circ \cdots \circ \sigma_{n-1}$$

for $n \ge k \ge 0$, where $\sigma_{[k,k)}$ is the identity map; then $\sigma_{[k,n)}$ is a substitution from A_n^* to A_k^*. The *language of $\boldsymbol{\sigma}$ at level k* is defined by

$$\mathcal{L}_{\boldsymbol{\sigma}}^{(k)} := \big\{w \in A_k^* \,:\, w \in \mathcal{F}(\sigma_{[k,n)}(A_n)) \text{ for infinitely many } n > k\big\},$$

and $\mathcal{L}_{\boldsymbol{\sigma}} := \mathcal{L}_{\boldsymbol{\sigma}}^{(0)}$. (In other papers, the requirement for infinitely many $n > k$ is replaced by "some $n > k$"; this can change the language only if a letter of A_m does not occur in σ_m.) Our definition ensures that

$$\mathcal{F}\big(\sigma_{[k,n)}(\mathcal{L}_{\boldsymbol{\sigma}}^{(n)})\big) = \mathcal{L}_{\boldsymbol{\sigma}}^{(k)} \quad \text{for all } n \ge k \ge 0. \tag{4.1}$$

A sequence of substitutions $(\sigma_k)_{k\ge 0}$ is *everywhere growing* if $\lim_{k\to\infty}\langle\sigma_{[0,k)}\rangle = \infty$. It is left (resp. right) *proper* when for each $k \ge 0$ there exists $n > k$ such that $\sigma_{[k,n)}$ is left (resp. right) proper. The following theorem was proved in [BCBD+21, Corollary 5.5] for unimodular incidence matrices, i.e., $|\det \mathcal{M}_{\sigma_k}| = 1$ for all $k \ge 0$.

Theorem 4.1. *Let $\boldsymbol{\sigma}$ be a left or right proper sequence of substitutions. If $\mathcal{L}_{\boldsymbol{\sigma}}^{(k)}$ is letter-balanced for infinitely many k, then $\mathcal{L}_{\boldsymbol{\sigma}}$ is factor-balanced.*

Proof. Assume that $\mathcal{L}_{\boldsymbol{\sigma}}^{(k)}$ is letter-balanced for infinitely many k, which implies that it is letter-balanced for all k by (4.1) and Proposition 3.2. Since $\boldsymbol{\sigma}$ is left or right proper, there exist $0 = k_0 < k_1 < k_2 < \cdots$ such that $\sigma_{[k_i,k_{i+1})}$ is left or right proper for all $i \geq 0$. Therefore, by Proposition 3.3, $\mathcal{L}_{\boldsymbol{\sigma}}$ is balanced for all lengths $n \geq 1$. □

The following corollary is the particular case of Theorem 4.1 with constant sequence $\boldsymbol{\sigma} = (\sigma, \sigma, \dots)$ for some substitution $\sigma : A^* \to A^*$; we write σ^∞ for $(\sigma, \sigma, \dots)$. The *language of a substitution* is $\mathcal{L}_\sigma := \mathcal{L}_{\sigma^\infty}$ (and consists of those $w \in A^*$ that are in $\mathcal{F}(\sigma^n(A))$ infinitely often).

Corollary 4.2. *Let $\sigma : A^* \to A^*$ be a substitution such that σ^k is left or right proper for some $k \geq 1$. If $\mathcal{L}_\sigma$ is letter-balanced, then $\mathcal{L}_\sigma$ is factor-balanced.*

For invertible incidence matrices, letter-balancedness at level 0 implies letter-balancedness at all levels by Proposition 3.5, which gives the following corollary of Theorem 4.1.

Corollary 4.3. *Let $\boldsymbol{\sigma} = (\sigma_k)_{k\geq 0}$ be a left or right proper sequence of substitutions with invertible incidence matrix $\mathcal{M}_{\sigma_k}$ for all $k \geq 0$. If $\mathcal{L}_{\boldsymbol{\sigma}}$ is letter-balanced, then $\mathcal{L}_{\boldsymbol{\sigma}}$ is factor-balanced.*

If $\boldsymbol{\sigma}$ is not proper, then we need balancedness for length 2 to infer factor-balancedness.

Theorem 4.4. *Let $\boldsymbol{\sigma}$ be an everywhere growing sequence of substitutions such that $\mathcal{L}_{\boldsymbol{\sigma}}^{(k)}$ is balanced for length 2 for infinitely many k. Then $\mathcal{L}_{\boldsymbol{\sigma}}$ is factor-balanced.*

Proof. By Proposition 3.3, balancedness of $\mathcal{L}_{\boldsymbol{\sigma}}^{(k)}$ for length 2 implies balancedness of $\mathcal{L}_{\boldsymbol{\sigma}}$ for length $\langle \sigma_{[0,k)} \rangle + 1$. Since $\boldsymbol{\sigma}$ is everywhere growing, this implies that $\mathcal{L}_{\boldsymbol{\sigma}}$ is factor-balanced. □

For primitive substitutions σ, a sufficient condition for balancedness for length n of $\mathcal{L}_\sigma$ is given in [Ada03, Theorem 22], and it indicates that balancedness for length 2 and factor-balancedness are closely related; see also [Que10, Section 5.4.3]. We prove that balancedness for length 2 implies factor-balancedness for most substitutions.

Corollary 4.5. *Let $\sigma : A^* \to A^*$ be an everywhere growing substitution. If $\mathcal{L}_\sigma$ is balanced for length 2, then $\mathcal{L}_\sigma$ is factor-balanced.*

We remark that, in Theorem 4.1, Corollary 4.3 and Theorem 4.4, factor-balancedness holds not only for $\mathcal{L}_{\boldsymbol{\sigma}}$ but for all $\mathcal{L}_{\boldsymbol{\sigma}}^{(k)}$, $k \geq 0$.

5. Thue–Morse–Sturmian languages

We conclude the paper by studying a special class of sequences of substitutions that occurs naturally in [Ste20, KSZ22]. A language $\mathcal{L}_{\boldsymbol{\sigma}}$, $\boldsymbol{\sigma} \in \{L, M, R\}^\infty$, with substitutions

$$\begin{aligned} L : 0 &\mapsto 0, & M : 0 &\mapsto 01, & R : 0 &\mapsto 01, \\ 1 &\mapsto 10, & 1 &\mapsto 10, & 1 &\mapsto 1, \end{aligned}$$

is *Thue–Morse–Sturmian* if $\boldsymbol{\sigma}$ is primitive. Recall that a sequence of substitutions $(\sigma_n)_{n\geq 0}$ is *primitive* if, for each $k \geq 0$, there exists $n > k$ such that $|\sigma_{[k,n)}(a)|_b \geq 1$ for all $a \in A_n$, $b \in B_k$; in the case of $\boldsymbol{\sigma} \in \{L, M, R\}^\infty$, this means that $\boldsymbol{\sigma}$ does not end with the constant sequence L^∞ or R^∞. However, the following results also hold for non-primitive sequences.

Proposition 5.1. *For all $\boldsymbol{\sigma} \in \{L, M, R\}^\infty$, $\mathcal{L}_{\boldsymbol{\sigma}}$ is letter-2-balanced.*

Proof. For $\boldsymbol{\sigma} = (\sigma_k)_{k\geq 0} \in \{L, M, R\}^\infty \setminus \{L, R\}^\infty$, let $n \geq 1$ be minimal such that $\sigma_n = M$. We claim that $\mathcal{F}(\sigma_{[0,n)}(\{01, 10\}^*))$ is letter-2-balanced. Indeed, we have $\sigma_{[0,n)}(01) = 01w$ and $\sigma_{[0,n)}(10) = 10w$ for some $w \in \{0, 1\}^*$. (This property is trivial for $\sigma_{[n,n)}$ and can be shown inductively for $\sigma_{[k,n)}$, $0 \leq k < n$, since $\sigma_k \in \{L, R\}$.) Let $v, v' \in \mathcal{F}(\sigma_{[0,n)}(\{01, 10\}^*))$ with $|v| = |v'|$. If $|v| \geq |w|+3$ then we can write $v = pus$, $v' = p'u's'$ such that $|u|_0 = |u'|_0 = |w|_0+1$, $|u|_1 = |u'|_1 = |w|_1+1$ and $ps, p's' \in \mathcal{F}(\sigma_{[0,n)}(\{01, 10\}^*))$. Since $|ps| = |p's'|$ and $|ps|_0 - |p's'|_0 = |v|_0 - |v'|_0$, it is sufficient to consider $|v| \leq |w|+2$. If $|v| \leq |w|+1$, then $v, v' \in \mathcal{F}(\sigma_{[0,n)}(\{01\}^*))$, and it is well known that this language is letter-1-balanced; see [Lot02, Chapter 2]. For $|v| = |w|+2$, we have $|v|_0 = |w|_0+1$ or $v \in \{0w0, 1w1\}$. Therefore, $\mathcal{F}(\sigma_{[0,n)}(\{01, 10\}^*))$ and thus $\boldsymbol{\mathcal{L}_\sigma}$ are letter-2-balanced.

Let now $\boldsymbol{\sigma} \in \{L, R\}^\infty$. If $\boldsymbol{\sigma}$ contains infinitely many L's and R's, then $\boldsymbol{\mathcal{L}_\sigma}$ is Sturmian and thus letter-1-balanced; see e.g. [Lot02, Chapter 2]. Since $L^n(0) = 0$, $L^n(1) = 10^n$, $R^n(0) = 01^n$, $R^n(1) = 1$, for all $n \geq 0$, the languages $\mathcal{L}_{L^\infty}$ and $\mathcal{L}_{R^\infty}$ are also letter-1-balanced. Finally, if $\sigma_n = L$, $\sigma_k = R$ for all $k > n$, or $\sigma_n = R$, $\sigma_k = L$ for all $k > n$, then $\boldsymbol{\mathcal{L}_\sigma} = \mathcal{F}(\sigma_{[0,n)}(\{01\}^*))$, which is again letter-1-balanced. □

For a characterisation of factor-balancedness of Thue–Morse–Sturmian languages, we need the following lemma in order to show that applying any composition of substitutions L, M, R to the Thue–Morse language, which is not factor-balanced, does not create a factor-balanced language. More precisely, we show for $\sigma \in \{L, M, R\}^*$ that $\sigma(011)$ occurs only trivially in $\sigma(w)$, $w \in \mathcal{L}_{M^\infty}$. Here, $\{L, M, R\}^*$ is the set of compositions of substitutions in L, M, R.

Lemma 5.2. *Let $\sigma \in \{L, M, R\}^*$, $a, b \in \{0, 1\}$, $p, s, v \in \{0, 1\}^*$, $k \geq 1$, such that*

$$\sigma(avb) = p\,\sigma(01^k)\,s,$$

p is a strict prefix of $\sigma(a)$ and s is a strict suffix of $\sigma(b)$. Then $avb = 01^k$ (and p, s are empty) or $avb = 1^{k+1}$ (and s is empty).

Proof. The statement is clearly true when σ is the identity. For $\sigma = \sigma_0 \circ \sigma_1 \circ \cdots \circ \sigma_n$, $\sigma_i \in \{L, M, R\}$, we prove the statement by induction on n.

Let first $\sigma_n = L$. Then we have

$$\sigma_{[0,n)}(L(avb)) = p\,\sigma_{[0,n)}(0(10)^k)\,s \quad \text{and} \quad 11 \notin \mathcal{F}(L(avb)).$$

If p is a strict prefix of $\sigma_{[0,n)}(a)$, in particular if $a = 0$, then $\sigma_{[0,n)}(av'b') = p\,\sigma_{[0,n)}(01)\,s'$ for a prefix $av'b'$ of $L(avb)$ and a strict suffix s' of $\sigma_{[0,n)}(b')$. By the induction hypothesis and since $11 \notin \mathcal{F}(L(av'b'))$, this implies that p is empty. Since $\sigma_{[0,n)}(0)$ starts with 0 and $\sigma_{[0,n)}(1)$ starts with 1, we obtain that $L(avb) = 0(10)^k$ and thus $avb = 01^k$. If $a = 1$ and $p = \sigma_{[0,n)}(1)\,p'$, then $\sigma_{[0,n)}(0v'b') = p'\sigma_{[0,n)}(01)\,s'$ for a prefix $10v'b'$ of $L(1vb)$ and a strict suffix s' of $\sigma_{[0,n)}(b')$. Now, the induction hypothesis implies that p' is empty, thus $L(avb) = (10)^{k+1}$, i.e., $avb = 1^{k+1}$.

Let now $\sigma_n = M$. Then we have

$$\sigma_{[0,n)}(M(avb)) = p\,\sigma_{[0,n)}(01(10)^k)\,s \quad \text{and} \quad 111 \notin \mathcal{F}(M(avb)).$$

If p is a strict prefix of $\sigma_{[0,n)}(a)$, then $\sigma_{[0,n)}(av'b') = p\,\sigma_{[0,n)}(011)\,s'$, hence p is empty, and $avb = 01^k$. If $a = 1$ and $p = \sigma_{[0,n)}(1)\,p'$, then we obtain that $avb = 1^{k+1}$. If $a = 0$ and $p = \sigma_{[0,n)}(0)\,p'$, then $\sigma_{[0,n)}(1v'b') = p'\sigma_{[0,n)}(011)\,s'$, hence p' is empty, which contradicts that $\sigma_{[0,n)}(1)$ and $\sigma_{[0,n)}(0)$ start with different letters.

Finally, let $\sigma_n = R$. Then we have

$$\sigma_{[0,n)}(R(avb)) = p\,\sigma_{[0,n)}(01^{k+1})\,s \quad \text{and} \quad 1^{k+2} \notin \mathcal{F}(R(avb)).$$

If p is a strict prefix of $\sigma_{[0,n)}(a)$, then $R(avb) = 01^{k+1}$, thus $avb = 01^n$. Otherwise, we have $a = 0$ and $p = \sigma_{[0,n)}(0)\,p'$, thus $\sigma_{[0,n)}(1v'b') = p'\sigma_{[0,n)}(01^{k+1})\,s'$, hence p' is empty, which contradicts that $\sigma_{[0,n)}(1)$ and $\sigma_{[0,n)}(0)$ start with different letters. □

Theorem 5.3. *Let $\boldsymbol{\sigma} = (\sigma_k)_{k\geq 0} \in \{L, M, R\}^\infty$. Then $\mathcal{L}_{\boldsymbol{\sigma}}$ is factor-balanced if and only if $\sigma_k \neq M$ for infinitely many k.*

Proof. By Proposition 5.1, $\mathcal{L}_{\boldsymbol{\sigma}}^{(k)}$ is letter-balanced for all $k \geq 0$. If $\boldsymbol{\sigma}$ does not end with M^∞, then it is right proper. If $\boldsymbol{\sigma}$ also does not end with L^∞ or R^∞, then it is everywhere growing, and we can apply Theorem 4.1. If $\boldsymbol{\sigma}$ ends with LR^∞ or RL^∞, then we have seen in the proof of Proposition 5.1 that $\mathcal{L}_{\boldsymbol{\sigma}} = \mathcal{F}(\sigma(\{01\}^*))$ for some $\sigma \in \{L, M, R\}^*$, which is factor-balanced. The cases of $\mathcal{L}_{L^\infty}$ and $\mathcal{L}_{R^\infty}$ are similar.

Consider now $\boldsymbol{\sigma}$ ending with M^∞. We first prove that the Thue–Morse language $\mathcal{L}_{M^\infty}$ is not balanced for length 2, giving more details than [Sad16, Example 3]; a more general proof can be found in [BCB19]. To this end, define recursively words $w_n, w'_n \in \mathcal{L}_{M^\infty}$ by

$$\begin{aligned} w_1 &= 00, \quad M^2(w_{2n-1}) = 0\,w_{2n}\,0, \quad M^2(w_{2n}) = 1\,w_{2n+1}\,1,\\ w'_1 &= 01, \quad M^2(w'_{2n-1}) = w'_{2n}\,01, \quad M^2(w'_{2n}) = w'_{2n+1}\,10, \end{aligned}$$

with $|w_n| = |w'_n| = \frac{4^n+2}{3}$. The second higher block codes are $(w_1)^{(2)} = (00)$, $(w'_1)^{(2)} = (01)$,

$$\begin{aligned} M_2^2((w_{2n-1})^{(2)})\,(01)(11) &= (01)\,(w_{2n})^{(2)}, \quad M_2^2((w_{2n})^{(2)})\,(10)(00) = (10)\,(w_{2n+1})^{(2)},\\ M_2^2((w'_{2n-1})^{(2)})\,(10) &= (w'_{2n})^{(2)}, \qquad\qquad M_2^2((w'_{2n})^{(2)})\,(01) = (w'_{2n+1})^{(2)}, \end{aligned}$$

with the substitution

$$M_2 : (\{0,1\}^2)^* \to (\{0,1\}^2)^*, \qquad \begin{aligned} &(00) \mapsto (01)(10), \quad (01) \mapsto (01)(11),\\ &(10) \mapsto (10)(00), \quad (11) \mapsto (10)(01). \end{aligned}$$

Using the abelianizations

$$\boldsymbol{\ell}_2(w) = \begin{pmatrix} |w|_{00}\\ |w|_{01}\\ |w|_{10}\\ |w|_{11} \end{pmatrix}, \quad \mathcal{M}_{M_2} = \big(|M_2(cd)|_{ab}\big)_{ab,cd\in\{00,01,10,11\}} = \begin{pmatrix} 0&0&1&0\\ 1&1&0&1\\ 1&0&1&1\\ 0&1&0&0 \end{pmatrix},$$

we obtain that

$$\begin{aligned} \boldsymbol{\ell}_2(w_{2n}) &= \mathcal{M}_{M_2}^2\,\boldsymbol{\ell}_2(w_{2n-1}) + \boldsymbol{\ell}_2(11), \quad \boldsymbol{\ell}_2(w_{2n+1}) = \mathcal{M}_{M_2}^2\,\boldsymbol{\ell}_2(w_{2n}) + \boldsymbol{\ell}_2(00),\\ \boldsymbol{\ell}_2(w'_{2n}) &= \mathcal{M}_{M_2}^2\,\boldsymbol{\ell}_2(w'_{2n-1}) + \boldsymbol{\ell}_2(10), \quad \boldsymbol{\ell}_2(w'_{2n+1}) = \mathcal{M}_{M_2}^2\,\boldsymbol{\ell}_2(w'_{2n}) + \boldsymbol{\ell}_2(00). \end{aligned}$$

The right eigenvectors of $\mathcal{M}_{M_2}$ (to the eigenvalues $2, -1, 0, 1$) are

$$\mathbf{v}_2 = \begin{pmatrix} 1\\2\\2\\1 \end{pmatrix}, \ \mathbf{v}_{-1} = \begin{pmatrix} 1\\-1\\-1\\1 \end{pmatrix}, \ \mathbf{v}_0 = \begin{pmatrix} 1\\0\\0\\-1 \end{pmatrix}, \ \mathbf{v}_1 = \begin{pmatrix} 1\\-1\\1\\-1 \end{pmatrix},$$

thus

$$\boldsymbol{\ell}_2(w_{2n}) = \frac{4^{2n}-1}{18}\mathbf{v}_2 + \frac{2n}{3}\mathbf{v}_{-1} - \frac{1}{2}\mathbf{v}_0 \quad \text{and} \quad \boldsymbol{\ell}_2(w'_{2n}) = \frac{4^{2n}-1}{18}\mathbf{v}_2 - \frac{n}{3}\mathbf{v}_{-1} - \frac{1}{2}\mathbf{v}_0,$$

hence $\boldsymbol{\ell}_2(w_{2n}) - \boldsymbol{\ell}_2(w'_{2n}) = n\,\mathbf{v}_{-1}$, i.e.,

$$|w_{2n}|_{00} - |w'_{2n}|_{00} = |w'_{2n}|_{01} - |w_{2n}|_{01} = |w'_{2n}|_{10} - |w_{2n}|_{10} = |w_{2n}|_{11} - |w'_{2n}|_{11} = n.$$

To finish the proof of the theorem, we have to show that $\mathcal{F}(\sigma(\mathcal{L}_{M^\infty}))$ is not factor-balanced for all $\sigma \in \{L, M, R\}^*$. Since $111 \notin \mathcal{L}_{M^\infty}$, Lemma 5.2 implies that $|\sigma(w)|_{\sigma(011)} = |w|_{011}$ for all $w \in \mathcal{L}_{M^\infty}$, and we clearly have $0 \le |w|_{11} - |w|_{011} \le 1$. Therefore, we have

$$\big||\sigma(w)|_{\sigma(011)} - |\sigma(w')|_{\sigma(011)}\big| \ge \big||w|_{11} - |w'|_{11}\big| - 1$$

for all $w, w' \in \mathcal{L}_{M^\infty}$. Since $|w|_{11} - |w'|_{11}$ is unbounded for $w, w' \in \mathcal{L}_{M^\infty}$ with $|w| = |w'|$, it is also unbounded when we restrict to w, w' with $|w|_0 = |w'|_0$ (and $|w|_1 = |w|'_1$). Then we have $|\sigma(w)| = |\sigma(w')|$, thus $\sigma(\mathcal{L}_{M^\infty})$ is not balanced for length $|\sigma(011)|$. □

We remark that, by Proposition 3.3, $\mathcal{F}(\sigma \circ L(\mathcal{L}_{M^\infty}))$ is balanced for length $|\sigma(0)| + 1$ for any substitution σ. On the other hand, we have seen in the proof of Theorem 5.3 that $\mathcal{F}(\sigma \circ L(\mathcal{L}_{M^\infty}))$ is not balanced for length $|\sigma(01010)|$ for any substitution $\sigma \in \{L, M, R\}^*$.

References

[Ada03] B. Adamczewski, *Balances for fixed points of primitive substitutions*, Theoret. Comput. Sci. **307** (2003), no. 1, 47–75.

[Ada04] ———, *Symbolic discrepancy and self-similar dynamics*, Ann. Inst. Fourier (Grenoble) **54** (2004), no. 7, 2201–2234 (2005).

[BCB19] V. Berthé and P. Cecchi Bernales, *Balancedness and coboundaries in symbolic systems*, Theoret. Comput. Sci. **777** (2019), 93–110.

[BCBD+21] V. Berthé, P. Cecchi Bernales, F. Durand, J. Leroy, D. Perrin, and S. Petite, *On the dimension group of unimodular $\mathcal{S}$-adic subshifts*, Monatsh. Math. **194** (2021), no. 4, 687–717.

[BT02] V. Berthé and R. Tijdeman, *Balance properties of multi-dimensional words*, Theoret. Comput. Sci. **273** (2002), no. 1-2, 197–224.

[FV02] I. Fagnot and L. Vuillon, *Generalized balances in Sturmian words*, Discrete Appl. Math. **121** (2002), no. 1-3, 83–101.

[KSZ22] V. Komornik, W. Steiner, and Y. Zou, *Unique double base expansions*, 2022, arXiv:2209.02373.

[Lot02] M. Lothaire, *Algebraic combinatorics on words*, Encyclopedia of Mathematics and its Applications, vol. 90, Cambridge University Press, Cambridge, 2002.

[MH40] M. Morse and G. A. Hedlund, *Symbolic dynamics II. Sturmian trajectories*, Amer. J. Math. **62** (1940), 1–42.

[Que10] M. Queffélec, *Substitution dynamical systems—spectral analysis*, second ed., Lecture Notes in Mathematics, vol. 1294, Springer-Verlag, Berlin, 2010.

[Sad16] L. Sadun, *Finitely balanced sequences and plasticity of 1-dimensional tilings*, Topology Appl. **205** (2016), 82–87.

[Ste20] W. Steiner, *Thue-Morse-Sturmian words and critical bases for ternary alphabets*, Bull. Soc. Math. France **148** (2020), no. 4, 597–611.

Université de Lyon, ENS de Lyon, Département informatique de l'ENS de Lyon, 46 allée d'Italie, F-69007 Lyon, France

Email address: leo.poirier@ens-lyon.fr

Université Paris Cité, CNRS, IRIF, F-75006 Paris, France

Email address: steiner@irif.fr