Theoretical Computer Science 934 (2022) 65–80

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Monochromatic arithmetic progressions in binary Thue–Morse-like words

Ibai Aedo [a,*], Uwe Grimm [a,1], Yasushi Nagai [b,1], Petra Staynova [c,*,2]

[a] School of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK
[b] School of General Education, Shinshu University, 3-1-1, Asahi, Matsumoto, Nagano, 390-8621, Japan
[c] School of Computing and Engineering, University of Derby, Kedleston Road, Derby DE22 1GB, UK

ARTICLE INFO

Article history:
Received 13 January 2021
Received in revised form 8 June 2022
Accepted 12 August 2022
Available online 19 August 2022
Communicated by M. Sciortino

Dedicated to our late friend and coauthor Uwe Grimm, who lived an inspiring life.

Keywords:
Combinatorics on words
Binary language
Infinite word
Thue–Morse sequence
Arithmetic progression
Bijective substitution

ABSTRACT

We study the length of monochromatic arithmetic progressions in the Thue–Morse word and in a class of generalised Thue–Morse words. In particular, we give exact values or upper bounds for the lengths of monochromatic arithmetic progressions of given fixed differences inside these words. Some arguments for these are inspired by van der Waerden's proof for the existence of arbitrary long monochromatic arithmetic progressions in any finite colouring of the (positive) integers. We also establish upper bounds for the length of monochromatic arithmetic progressions of certain differences in any fixed point of a primitive binary bijective substitution.

1. Introduction

We study the Thue–Morse word, also known as the Prouhet–Thue–Morse word. We refer the reader to [3,5], which is, respectively contains, an extensive survey of its properties. There are many alternative ways to define this word, one being through the sum of digits in the binary expansion of the integers indexing the word. More precisely, for each non-negative integer n, we determine the nth term of the Thue–Morse word by summing all the digits in the binary expansion of n and taking this sum modulo 2. The first few terms are given below:

$$01101001100101101001\cdots .$$

In this paper, we are interested in monochromatic arithmetic progressions that we can find in the Thue–Morse word, as well as in other generalized Thue–Morse words. Let us first make the notion of monochromatic arithmetic progression precise. For any positive integer c, a *c-colouring* of a set S of integers is a map from S to a set of c distinct colours, and for

* Corresponding authors.
E-mail addresses: ibai.aedo@open.ac.uk (I. Aedo), petra.staynova@gmail.com (P. Staynova).
[1] Supported by the Engineering and Physical Sciences Research Council (EPSRC) via Grant EP/S010335/1.
[2] Supported by an Early Career Fellowship from the London Mathematical Society.

https://doi.org/10.1016/j.tcs.2022.08.013

any positive integers d and L, an *arithmetic progression* of difference d and length L is a sequence $n, n+d, n+2d, \ldots, n+(L-1)d$ of integers. Then, a *monochromatic arithmetic progression* is an arithmetic progression all of whose elements have been assigned the same colour.

Our results draw inspiration from one of the best-known Ramsey-type theorems, namely, van der Waerden's theorem. In his seminal paper [36], van der Waerden showed that, in any colouring of the integers, the lengths of monochromatic arithmetic progressions of any difference cannot be bounded by a constant. More precisely,

Theorem 1 *(van der Waerden [36]). Let L and c be positive integers. There exists a positive integer N such that any c-colouring of the segment $\{0, 1, \ldots, N-1\}$ contains a monochromatic arithmetic progression of length L.*

It follows from van der Waerden's theorem that, given an infinite sequence v over a finite alphabet, v contains monochromatic arithmetic progressions of every positive integer length L. It is then natural to ask what happens if the difference d of the arithmetic progressions is fixed: do there exist arbitrarily long monochromatic arithmetic progressions of difference d in v? and do there exist infinite ones? The subtle difference between these two questions can be easily illustrated by the following examples. The binary sequence $v = 0^{10}1^{10}0^{100}1^{100}0^{1000}1^{1000}\cdots$ contains arbitrarily long finite monochromatic arithmetic progressions but does not contain infinite ones. The Champernowne word, obtained by concatenating the decimal representation of the positive integers, would be another example. The non-existence of infinite monochromatic arithmetic progressions within infinite words has been addressed in [25,37].

It is known that for some infinite words there are no arbitrarily long finite monochromatic arithmetic progressions if we fix the difference. Morgenbesser et al. [27] showed this in the case of the Thue–Morse sequence, and Parshina [31,33] studied the largest monochromatic arithmetic progressions of a fixed difference that it contains.

The problem of what types of finite words can occur as arithmetic progressions within substitution (and other) sequences has been considered in several contexts. In particular, the notion of arithmetic complexity, which generalises the subword complexity, has been studied in [6,12,17,18]. Another interesting result [7] states that any binary word appears as an arithmetic progression within the Thue–Morse word, and other related questions on Thue–Morse type systems concern Gowers uniformity norms [23] or prefix palindromic lengths [19].

This paper is organised as follows. In Section 2, we give background material on symbolic substitutions and symbolic dynamical systems. In Section 3, we prove the non-existence of arbitrarily long monochromatic arithmetic progressions in a family of infinite words (Proposition 8) a member of which is the Thue–Morse word, thus generalizing the result in [27]. We continue in Section 4 by setting out our main technique. This allows us, in Theorem 15, to consider arithmetic progressions in the Thue–Morse word and re-establish a result of Parshina [31] in a different way. In Theorem 21, we extend this for other differences. In Section 5, we consider other binary words with a similar substitution structure [9,10,22,26]. In Theorems 32 and 33, we apply our techniques to generalise some of our previous results, and we state Conjecture 35 for certain progression differences. We give illustrative plots at the end of sections 4 and 5. In Section 6, we establish upper bounds for the length of monochromatic arithmetic progressions of certain differences in any fixed point of a primitive binary bijective substitution (Propositions 39 and 40). We conclude in Section 7, with some open questions and proposed directions for further research.

2. Preliminaries

We use standard texts [11,24] for the theory of combinatorics on words and substitutions. We use $\mathbb{N}$ to denote the set of non-negative integers and $\mathbb{N}^+$ for the set of positive integers. In general, an *alphabet* is a non-empty finite subset $\mathcal{A}$ of $\mathbb{R}$, and its elements are called *letters*. Throughout this paper, we will work with the binary alphabet $\mathcal{A} = \{0, 1\}$.

Letters can be concatenated into *words*; we denote by $\mathcal{A}^+$ the set of non-empty finite words over an alphabet $\mathcal{A}$, by $\mathcal{A}^{\mathbb{N}}$ the set of (right-) infinite words over $\mathcal{A}$, and by $\mathcal{A}^{\mathbb{Z}}$ the set of bi-infinite words over $\mathcal{A}$. For a finite word $w = w_0 w_1 \ldots w_{n-1} \in \mathcal{A}^+$, we denote its length by $|w| = n$. For any finite or infinite word w, we denote by w_i its letter at position $i < |w|$, and by $w_{[i,j)}$ its *subword* (factor) $w_i w_{i+1} \ldots w_{j-1}$ of length $j - i$, for $0 \leqslant i < j \leqslant |w|$, with analogous definitions for bi-infinite words. In the following sections, we will use the terms 'word' and 'sequence' interchangeably when the context is clear.

A *substitution* is a map $\phi : \mathcal{A} \to \mathcal{A}^+$; this can be extended to a map on $\mathcal{A}^+$, $\mathcal{A}^{\mathbb{N}}$, $\mathcal{A}^{\mathbb{Z}}$ via concatenation. We say a substitution has *constant length* if $|\phi(a)| = |\phi(b)|$ for all $a, b \in \mathcal{A}$. To avoid degeneracy, we consider primitive substitutions, in other words ones where there exists a power n such that for any $a, b \in \mathcal{A}$ we have that a occurs in $\phi^n(b)$. If for a letter $a \in \mathcal{A}$ we have that $\phi(a)$ begins with a, we have that the primitive substitution has a fixed right-infinite point. Given a substitution ϕ, we may define its *language* as the set $\mathcal{L}_\phi := \{w \in \mathcal{A}^+ : \exists n \in \mathbb{N} \text{ s.t. } w \text{ is a subword of } \phi^n(a)\}$. It can be seen that the language of a primitive substitution is independent of choice of letter a. We call the words in $\mathcal{L}_\phi$ *legal* or *ϕ-legal*.

The Thue–Morse substitution is defined on the binary alphabet $\mathcal{A} = \{0, 1\}$ as $\phi(0) = 01,\ \phi(1) = 10$. The fixed point of this substitution starting with the letter 0 gives rise to the Thue–Morse sequence. We have noted before that this sequence can be generated via the sum of digits modulo 2. More generally, we have that a binary bijective constant-length substitution can be represented via the sum of digits modulo 2 in base k, where k is the length of the substitution [4]. There are several natural ways to generalise the Thue–Morse sequence. One approach, taken by Parshina in [31], is to consider the so-called cyclic substitution on n letters. For an alphabet $\mathcal{A} = \{0, \ldots, n-1\}$, we define the cyclic Thue–Morse substitution as

$\phi(0) = 01\cdots(n-1)$, $\phi(1) = 12\cdots 0, \ldots,$ $\phi(n-1) = (n-1)01\ldots(n-2)$. Another natural way to generalise the Thue–Morse sequence is by considering other binary sequences with a similar substitution structure [9,10,22,26]. It is this class that is studied further in Section 5.

In the following section it will be useful to consider the dynamical system associated with a substitution. To do this, we consider the discrete topology on the alphabet $\mathcal{A}$, which naturally induces the product topology on $\mathcal{A}^+$, $\mathcal{A}^{\mathbb{N}}$, and $\mathcal{A}^{\mathbb{Z}}$. The latter are compact by Tychonoff's theorem. We equip $\mathcal{A}^{\mathbb{N}}$ (respectively $\mathcal{A}^{\mathbb{Z}}$) with the (left) shift operator T, which gives these sets the structure of dynamical systems. A substitution gives rise to a natural dynamical system via the shift-orbit closure of its fixed points. More precisely, given a fixed point w of a primitive substitution ϕ, we define the set $X_\phi := \overline{\{T^n(w) : n \in \mathbb{N}\}} \subset \mathcal{A}^{\mathbb{N}}$, which is also known as the *hull* or the *subshift* generated by ϕ. The choice of fixed point is irrelevant if the substitution is primitive. By a simple compactness argument, the subshift of a substitution is precisely the set of elements of $\mathcal{A}^{\mathbb{N}}$ (resp $\mathcal{A}^{\mathbb{Z}}$) that have only ϕ-legal factors.

3. Pure point spectrum and infinite monochromatic arithmetic progressions

If v is a fixed point of a primitive constant-length substitution, the existence of arbitrarily long monochromatic arithmetic progressions is related to the spectral theory of the corresponding dynamical system. Note that we can alternatively consider a bi-infinite sequence $w \in \mathcal{A}^{\mathbb{Z}}$ that is a repetitive fixed point of the same substitution and ask whether for every $m \in \mathbb{N}^+$, there exists $n \in \mathbb{Z}$ such that $w_n = w_{n+id}$, for $0 \leqslant i \leqslant m-1$. This problem on bi-infinite sequences w is equivalent to the original problem on v, since both define the same language and hence the same shift space.

We recall the fact that for a constant-length substitution, its dynamical system (with $\mathbb{Z}$ shift-action) has pure point (discrete) spectrum if and only if the corresponding tiling dynamical system (with $\mathbb{R}$ translation-action) has pure point spectrum. Using this fact and [29, Theorem 5.1], we have the following.

Theorem 2. *Let w be a bi-infinite fixed point of a primitive, constant-length substitution ϱ. Then w contains infinite monochromatic arithmetic progressions if and only if ϱ has pure point dynamical spectrum.* □

While having pure point spectrum is a measure-theoretic property, for primitive constant-length substitutions this is reduced to an algorithmic check by the seminal result by Dekking [14]. Before we proceed, we will need some definitions. Throughout, we assume ϱ to be a primitive constant-length substitution of length L over a finite alphabet of r letters, and v to be a one-sided fixed point of ϱ.

Definition 3. The *height* $h(\varrho)$ of ϱ is given by

$$h(\varrho) = \max\{n \geqslant 1 \mid \gcd(n, L) = 1,\ n \text{ divides } \gcd\{a \mid v_0 = v_a\}\}.$$

Remark 4. The height satisfies $1 \leqslant h(\varrho) \leqslant r$. There is an algorithm to compute $h(\varrho)$ for a given substitution ϱ; see [14].

Definition 5. Let $j \in \{0, \ldots, L-1\}$. The *column* ϱ_j at position j is the map $\varrho_j \colon \mathcal{A} \to \mathcal{A}$ defined via $\varrho_j(a) := \varrho(a)_j$ for $a \in \mathcal{A}$. A column ϱ_j is called a *coincidence* if ϱ_j sends all letters to a single image, i.e., for all $a \in \mathcal{A}$, $\varrho_j(a) = b$ for some fixed $b \in \mathcal{A}$.

Theorem 6 *([14, Theorem 7]). Let ϱ be a primitive constant-length substitution of height* 1. *Then ϱ has pure point spectrum if and only if it has a coincidence.* □

It was also shown in [14] that for every substitution ϱ, there exists a substitution ϱ' with $h(\varrho') = 1$ (called the pure base of ϱ) such that ϱ has pure point spectrum if and only if ϱ' has pure point spectrum. The pure base ϱ' can be directly derived from ϱ. In particular, if $h(\varrho) = 1$ then it is its own pure base.

Example 7. Let $\mathcal{A} = \{0, 1\}$ and consider the period doubling substitution

$$\varrho_{\mathrm{pd}} \colon \begin{array}{l} 0 \mapsto 01 \\ 1 \mapsto 00. \end{array}$$

One can easily verify that this has height one and since it has a coincidence column at the zeroth position, it has pure point spectrum by Theorem 6. In fact, the result in Theorem 2 is even more explicit because the fixed point v of ϱ_{pd} is a Toeplitz sequence, i.e., for every $n \in \mathbb{N}$, there exists $p_n \in \mathbb{N}$ such that $v_n = v_{n+p_n m}$ for all $m \in \mathbb{N}$; see [18,21]. In other words, every letter v_n is part of an infinite monochromatic arithmetic progression.

We now have the following result on arbitrarily long monochromatic progressions; compare [29, Lem. 2.5].

Proposition 8. *Let ϱ be a primitive constant-length substitution whose pure base does not have a coincidence. Then any of its fixed points does not admit arbitrarily long monochromatic arithmetic progressions of any difference d.*

Proof. The proof of this proposition uses the compactness of the corresponding subshift X_ϱ generated by the substitution ϱ. It follows from this that every sequence $\{x_n\}_{n\geqslant 0}$ with $x_n \in X_\varrho$ admits a converging subsequence.

Now pick a substitution which is primitive and whose pure base does not have a coincidence. It follows from Dekking's result that it must have mixed spectrum and hence by Theorem 2 it cannot have infinite monochromatic progressions.

We then assume that some fixed point v of ϱ admits arbitrarily long monochromatic progressions of some difference d. Then, we can find positions i_j for all $j \in \mathbb{N}$ such that v_{i_j} is the first letter of a finite arithmetic progression of a fixed colour, difference d and length L_j, satisfying $L_k > L_j$ for $k > j$, meaning that the lengths of the progressions are increasing.

Now let T be the left shift map defined pointwise via $(Tv)_i = v_{i+1}$. The sequence $\left\{T^{i_j} v\right\}_{j\geqslant 0}$ of shifted words admits a convergent subsequence and any limit word $u = \lim_{j_k\to\infty} T^{i_{j_k}} v$ must contain an infinite monochromatic arithmetic progression of difference d starting with its first letter. This contradicts Theorem 2 and completes the proof. □

We then recover the following well-known fact regarding the Thue–Morse sequence; see [27], which is based on [20]. Note that the Thue–Morse substitution is its own pure base and it does not have a coincidence.

Fact 9. *The Thue–Morse word does not contain arbitrarily long monochromatic arithmetic progressions of any fixed difference d.* □

Remark 10. Note that Proposition 8 allows one to do the same analysis on larger classes of automatic sequences not covered by [27], e.g., those which are codings of fixed points of constant-length substitutions which are a priori known to have mixed spectrum. This is addressed in an ongoing work [2].

For the rest of the paper, we will be dealing with infinite words generated by substitutions which satisfy the conditions of Proposition 8. For these objects we introduce the following well-defined notion, which we adapt from [31,32].

Definition 11. Let v be a fixed point of a substitution over an alphabet satisfying the conditions of Proposition 8. For a positive integer d, we denote by $A_v(d)$ the maximum length of a monochromatic arithmetic progression of difference d within the word v.

When it is clear from context, we drop the v in $A_v(d)$ and just refer to it as $A(d)$.

4. Monochromatic arithmetic progressions in the Thue–Morse word

We consider the Thue–Morse word $v \in \{0,1\}^{\mathbb{N}}$ arising from the substitution

$$\theta\colon \begin{array}{l} 0 \mapsto 01 \\ 1 \mapsto 10, \end{array} \tag{1}$$

as the fixed point $v = \lim_{n\to\infty} \theta^n(0)$. Note that this substitution is bijective and symmetric under the 'bar' operation that exchanges the two letters (so $\overline{a} = 1 - a$ for $a \in \{0,1\}$, referred to as a 'bar-swap symmetry' in [8]), which also implies that $\overline{v} = \lim_{n\to\infty} \theta^n(1)$ is another fixed point word. The word $v = v_0 v_1 v_2 \ldots$ satisfies

$$v_{2i} = v_i \quad \text{and} \quad v_{2i+1} = \overline{v_i}\,,$$

for all $i \in \mathbb{N}$. The letter v_i is thus 0 if the binary expansion of i contains an even number of 1 s, and 1 otherwise. Also, v is overlap-free, which means that, for any finite, non-empty word w, v does not contain www_0 as a subword, where w_0 denotes the first letter of w. Note also that $\theta^n(a)$ is reflection-symmetric if n is even, and antisymmetric (meaning that the reflected word is the image under the bar operation) if n is odd.

For $n \in \mathbb{N}$, we have the well-known recursions

$$\theta^{n+1}(a) = \theta^n(a)\,\overline{\theta^n(a)} = \theta^n(a)\,\theta^n(\overline{a}), \tag{2}$$

as can easily be shown by induction. This implies the following property.

Lemma 12. *For all $m > n \in \mathbb{N}^+$ and $a \in \{0,1\}$, the word $\theta^m(a)$ consists of a sequence of the two subwords $w = \theta^n(a)$ and $\overline{w} = \theta^n(\overline{a})$, arranged according to the sequence that corresponds to $\theta^{m-n}(b)$ with the letters b and $\overline{b}$ replaced by the words w and $\overline{w}$.*

Proof. Let $w = \theta^n(a)$ and $\overline{w} = \theta^n(\overline{a})$. Then, $\theta(w) = w\overline{w}$ by the recursion (2), which is the same form as the Thue–Morse substitution, now on the alphabet $\{w, \overline{w}\}$. Hence $\theta^m(a) = \theta^{m-n}(\theta^n(a)) = \theta^{m-n}(w)$ is the sequence $\theta^{m-n}(b)$ with $b, \overline{b}$ replaced by $w, \overline{w}$. □

Since $\theta(v) = v$, we know that $A(2^n d) = A(d)$ for any $n \in \mathbb{N}$. In particular, since v is overlap-free, this implies $A(2^n) = A(1) = 2$ for all $n \in \mathbb{N}$. The following results show that $A(d) = 2$ holds only for differences d that are powers of 2.

Lemma 13. *Let $d > 1$ be an odd integer. Then $A(d) \geqslant 3$.*

Proof. Assume first that the binary expansion of d contains an even number of 1 s. Since multiplication by 2 conserves the number of 1 s, we have $v_0 = v_d = v_{2d} = 0$ and so $A(d) \geqslant 3$.

Now consider the case that the binary expansion of d contains an odd number of 1 s, and hence at least 3. Write $d = 2^m + 2^n + k$, with $m > n$ and $k < 2^n$, so k again contains an odd number of 1 s in its binary expansion. Let $i = 2^{m+1} + 2^n$, with $v_i = 0$. Then $i + d = 2^{m+1} + 2^m + 2^{n+1} + k$. If $m > n+1$, the number of 1 s in the binary expansion of $i + d$ is even. If $m = n+1$, then $i + d = 2^{n+3} + k$ and so, the number of 1 s in its binary expansion is even too. Hence, $v_{i+d} = 0$. Furthermore, $i + 2d = 2^{m+2} + 2^{n+1} + 2^n + 2k$. If $2k < 2^n$, the number of 1 s in the binary expansion of $i + 2d$ is even. If $2^n \leqslant 2k < 2^{n+1}$, we write $2k = 2^n + t$, where the number of 1 s in the binary expansion of t is even, and so, $i + 2d = 2^{m+2} + 2^{n+2} + t$ and its binary expansion has an even number of 1 s. Therefore, $v_{i+2d} = 0$ and $A(d) \geqslant 3$. □

Corollary 14. *$A(d) = 2$ if, and only if, $d = 2^n$ for some $n \in \mathbb{N}$.*

Proof. Since $A(2^n d) = A(d)$ for all $n \in \mathbb{N}$ and $A(d) \geqslant 3$ for all odd $d > 1$ by Lemma 13, $A(d) = 2$ implies that d contains no odd prime factors. □

Parshina proved the following result [31], as well as a generalisation to similar sequences in larger alphabets [32,33]. Her proofs for the Thue–Morse case [31] are based on a detailed analysis of binary arithmetic.

Theorem 15 *([31]). For all $n \in \mathbb{N}^+$, we have that*

$$\max_{d<2^n} A(d) = A(2^n - 1) = \begin{cases} 2^n + 4, & \text{if } 2 \mid n, \\ 2^n, & \text{otherwise.} \end{cases}$$

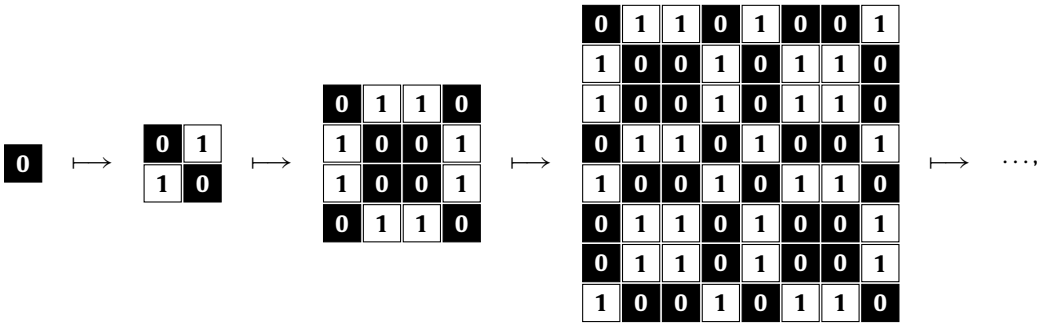
We will give exact expressions of $A(d)$ for certain values of d, in this section for the Thue–Morse sequence, and in Section 5 also for generalised Thue–Morse sequences, which are different generalisations from those considered in [31–33]. In particular, all our sequences are on a two-letter alphabet. Our results include a simple proof of the value of $A(2^n - 1)$ stated in Theorem 15, but also identify a second series of long monochromatic arithmetic progressions in the Thue–Morse word, which becomes the 'longest' in some cases, provided one considers the maximum over a different range for d.

Our proofs will follow van der Waerden's argument. For this, let us define the following block substitution.

Definition 16. Let Θ be the block substitution on the alphabet $\{0, 1\}$ defined by

$$\Theta\colon \quad 0 \longmapsto \begin{matrix} 0\,1 \\ 1\,0, \end{matrix} \quad 1 \longmapsto \begin{matrix} 1\,0 \\ 0\,1. \end{matrix}$$

Iterating Θ on a single letter produces square blocks of size $2^n \times 2^n$, for instance



where we used black (for 0) and white (for 1) squares to emphasise the block structure. Note that the blocks along both diagonals are always of the same colour.

Lemma 17. *For $a \in \{0, 1\}$ and $n \in \mathbb{N}^+$, the block $\Theta^n(a)$, read row-wise from top to bottom, is the word $\theta^{2n}(a)$ with θ the Thue–Morse substitution of Equation (1).*

Proof. This follows by induction from noticing that

$$\Theta\colon \quad 0 \mapsto \begin{matrix} \theta(0) \\ \theta(1), \end{matrix} \quad 1 \longmapsto \begin{matrix} \theta(1) \\ \theta(0), \end{matrix}$$

so that, read row-wise from the top, the image of a under Θ is $\theta(a)\theta(\overline{a}) = \theta^2(a)$. □

The images of letters under Θ^n have the following properties.

Lemma 18. *For $a \in \{0, 1\}$ and $n \in \mathbb{N}^+$, the blocks $\Theta^n(a)$ consists of only two types of row and column words, and are symmetric under reflection in either diagonals. All entries on the main diagonal are a, while entries of the other diagonal are a for even n and $\overline{a}$ otherwise.*

Proof. As shown in Lemma 17, $\Theta^n(a)$ when read row-wise from the top is the word $\theta^{2n}(a) = \theta^n(\theta^n(a))$. By Lemma 12, this consists of 2^n words from $\{\theta^n(0), \theta^n(1)\}$. So each row is one of these two words.

The symmetry in the diagonals follows from the symmetry of the block substitution Θ, which also implies that all columns are either $\theta^n(0)$ or $\theta^n(1)$.

It is obvious from the inflation rule that the elements of $\Theta^n(a)$ on the main diagonal are always a. On the other diagonal, note that $\Theta(a)$ has $\overline{a}$ while $\Theta^2(a)$ has a, which implies the claim. □

Lemma 19. *For all $n \in \mathbb{N}^+$, we have that $A(2^n \pm 1) \geqslant 2^n$. For even n, we further have $A(2^n - 1) \geqslant 2^n + 2$.*

Proof. Consider the block $\Theta^n(a)$, which, when read row-wise from the top, is the word $\theta^{2n}(a)$. As shown in Lemma 17, all elements on the main diagonal are a, so we find that $A(2^n + 1) \geqslant 2^n$. Similarly, the elements on the other diagonal are either all a or $\overline{a}$, so we also have $A(2^n - 1) \geqslant 2^n$. For even n, both diagonals have a entries, and so the first and last letter of $\theta^{2n}(a)$ are also part of the arithmetic progression of difference $2^n - 1$, which implies the claim. □

Before we establish the values for $A(2^n \pm 1)$, we prove a useful result, which exploits the recognisability of the substitution; see [11] and references therein for general background.

Lemma 20. *For $n > 1$ and $a \in \{0, 1\}$, the word $w = \theta^n(a)$ occurs in the Thue–Morse word either as the level-n superword itself, or in the centre of two level-n superwords $\theta^n(\overline{a})\,\theta^n(\overline{a})$. Furthermore, if w is followed by the letter $\overline{a}$, or if w is preceded by the letter a (for n odd) or $\overline{a}$ (for n even), it is the level-n superword.*

Proof. Clearly w can occur as the level-n superword. The second possibility arises from

$$\begin{aligned}\theta^n(\overline{a}a) &= \theta^n(\overline{a})\,\theta^n(\overline{a}) \\ &= \theta^{n-1}(\overline{a}a)\,\theta^{n-1}(\overline{a}a) \\ &= \theta^{n-1}(\overline{a})\,\theta^{n-1}(a)\,\theta^{n-1}(\overline{a})\theta^{n-1}(a) \\ &= \theta^{n-1}(\overline{a})\,w\,\theta^{n-1}(a).\end{aligned}$$

To show that these are the only two possibilities, we use that

$$\theta^n(a) = \theta^{n-2}(a\overline{a}\overline{a}a) = \theta^{n-2}(a)\,\theta^{n-2}(\overline{a})\,\theta^{n-2}(\overline{a})\,\theta^{n-2}(a),$$

which holds for all $n > 1$. By recognisability, the two adjacent level-$(n-2)$ superwords $\theta^{n-2}(\overline{a})$ cannot belong to the same level-$(n-1)$ superword, so we know that $\theta^n(a)$ has to consist of two level-$(n-1)$ superwords, which only leaves the two possibilities, since all level-$(n-1)$ boundaries are determined.

If $w = \theta^{n-1}(a)\,\theta^{n-1}(\overline{a})$ is followed by a letter $\overline{a}$, the next level-$(n-1)$ superword is determined to be $\theta^{n-1}(\overline{a})$, and the level-$n$ superword boundary has to fall between w and the subsequent letter a, which shows that w is the level-n superword. The same happens when w is preceded by the final letter of the superword $\theta^{n-1}(a)$, which is a for odd n and $\overline{a}$ for even n. □

Theorem 21. *For all $n > 1$, we have that $A(2^n + 1) = 2^n + 2$.*

Proof. We first show that there exist monochromatic arithmetic progressions of length $2^n + 2$. From the proof of Lemma 19, we already have a monochromatic arithmetic progression of length 2^n in the word $w = \theta^{2n}(a)$, with the first and final letter being part of the progression.

Now consider how many letters can be added at either end of the progression of length 2^n in the superword $w = \theta^{2n}(a)$. From Lemma 20, we know that the word $w = \theta^{2n-1}(a)\,\theta^{2n-1}(\overline{a})$ and that these are the actual level-$(2n-1)$ superwords. There are four possibilities how this superword can be bordered by level-$(2n-1)$ superwords: we can have $\theta^{2n-1}(b)\,w\,\theta^{2n-1}(c)$ with $b, c \in \{a, \overline{a}\}$.

Since $d = 2^n + 1$, no element of the progression is in the level-n superwords adjacent at either end. Since all superwords start or end with a level-2 superword $b\overline{b}\overline{b}b$, the next two members on either side would have to be the first and the second

letter of the same superword $\theta^n(\overline{b})$, which however are different letters (where we use that $n > 1$). This shows that the progression can at most be extended by one in either direction. Since all combinations of superwords on either side can appear, there are instances where the progression can be extended by exactly one step in both directions, showing that $A(2^n{+}1) \geq 2^n + 2$.

It remains to be shown that this is the maximum length of a progression. Assume that we have a progression of length $L > 2^n$. The elements in this progression hit each position in the superwords of level-n at least once. Now, once we hit the first position of such a superword, the following members of the progression determine the sequence of level-n superwords uniquely, which is the same sequence as that of the superword w. If there are at least 2^n terms in the progression following this position, they determine the level-$(2^n{-}1)$ superword by the second part of Lemma 20, and hence we are back considering the word w from above. If there are fewer terms left, we can use the previous member of the progression which hits in the last position of a level-n superword, and determine the sequence of level-n superwords preceding it in the progression. Again, this determines the level-$(2^n{-}1)$ superword by the second part of Lemma 20, and we are back in the case considered above, showing that $L \leq 2^n + 2$. □

Similarly, as mentioned above, we can rederive the value of $A(2^n{-}1)$ stated in Theorem 15.

Proposition 22. *For all $n \in \mathbb{N}$, $n > 1$, we have that*

$$A(2^n{-}1) = \begin{cases} 2^n + 4, & \text{if } 2|n, \\ 2^n, & \text{otherwise.} \end{cases}$$

Proof. From Lemma 19, we already know that $A(2^n{-}1) \geqslant 2^n$ for n odd and $A(2^n{-}1) \geqslant 2^n + 2$ for n even.

Let us first consider the case that n is odd. Since the superwords $\theta^n(a)$ are antisymmetric under reflection, their first and last letters differ. This means that once our progression hits the first letter of a superword, it stops. Since addition by $2^n - 1$ means that the elements in the progression cycle through all positions in the superword, we obtain the upper limit $A(2^n{-}1) \leqslant 2^n$, implying that $A(2^n{-}1) = 2^n$.

Now consider the case of n even. Here, the superwords are symmetric under reflection, so we can have two elements of the progression within one superword. Assume that this occurs for a superword $\theta^n(a)$ which has first and last letter a. If the progression continued to the left and to the right, the neighbouring superwords are determined by having the letter a at the next two positions, which force both of them to be $\overline{a}$, and by symmetry this applies to either side of the superword, hence we obtain $\theta^n(\overline{a})\theta^n(\overline{a})\theta^n(a)\theta^n(\overline{a})\theta^n(\overline{a})$. Clearly, the word $\overline{a}\overline{a}a\overline{a}\overline{a}$ does not belong to the Thue–Morse language. This means that once the progression hits the first and last letter, it can only be extended by at most one either way, so we obtain $A(2^n{-}1) \leqslant 2^n + 4$. That this bound is attained can be seen by looking at $\theta^{2n}(a)$, which by Lemma 19 contains a progression of length $2^n + 2$ starting at ending with a superword $\theta^n(a)$ that contain two elements of the progression. The superwords either side of $\theta^{2n}(a)$ can be both $\theta^{2n}(\overline{a})$, since $\overline{a}a\overline{a}$ is in the Thue–Morse language. Hence the progression can be continued by one additional step to either direction, and the bound is attained. □

The following two lemmas prove that there are no longer monochromatic arithmetic progressions for differences up to powers of 2.

Lemma 23. *Let $n \in \mathbb{N}^+$ and $0 < k < 2^{n-1}$ be both odd, and consider $d = 2^n - k$. Then $A(d) \leqslant 2^n$.*

Proof. Assume that there exists a monochromatic arithmetic progression of difference d of length $L > 2^n$. Since d is odd and hence coprime with 2^n, looking at elements of the progression within superwords of length 2^n will meet every position in a superword, including the position $m = (k-1)/2$. However, the superwords of length 2^n for n odd are antisymmetric under reflection, so

$$\theta^n(a)_m = \overline{\theta^n(a)_{2^n-1-m}},$$

which shows that not both m and $2^n - 1 - m = m + d$ can be in the arithmetic progression, in contradiction to our assumption. This means that the progression cannot be longer than the number of rest classes modulo 2^n, which establishes the claim. □

Lemma 24. *Let $n > 1$ and let $2 < k < 2^{n-1}$ be odd, and consider $d = 2^n - k$. Then $A(d) \leqslant 2^n$.*

Proof. Assume there exists a monochromatic arithmetic progression of difference d of length $L > 2^n$. Since d is odd and hence coprime with 2^n, looking at elements of the progression within superwords of length 2^n will meet every position in a superword. There are precisely k instances where two elements of the progression appear within the same the superwords of length 2^n, and hence the corresponding letters within the superwords have to agree. Since $\theta^n(\overline{a}) = \overline{\theta^n(a)}$, both superwords

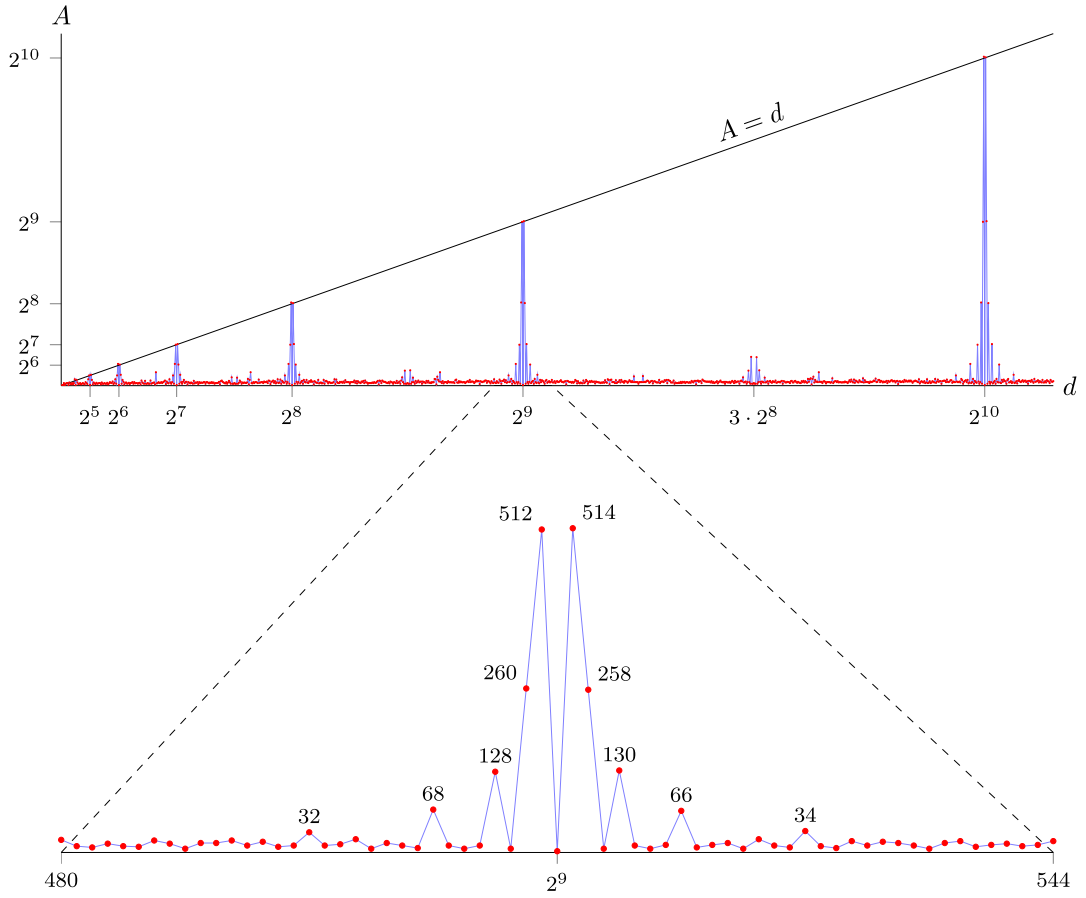


Fig. 1. $A(d)$ for $d = 1, 2, \ldots, 1100$ in the Thue–Morse word.

have to agree on all these positions. As a consequence, for $a \in \{0, 1\}$ and the superword $\theta^n(a)$, the word consisting of its first k letters

$$w := \theta^n(a)_{[0,k)}$$

also has to appear at the end of the superword, so $w = \theta^n(a)_{[2^n-k,2^n)}$. Since $k \geqslant 3$ and $\theta^n(a)$ starts with $a\overline{aa}$, the word w always contains a repeated letter and hence the level-1 superwords of length 2 are uniquely determined. This results in a contradiction because the length of w is odd, and the superword $\theta^n(a)$ thus cannot end in w, since the level-1 superword boundaries do not match. Hence $A(d) \leqslant 2^n$. □

Note that, in contrast to Lemma 23, the result of Lemma 24 does not extend to the case $k = 1$, since in this case there is only one instance of a word containing two elements of the arithmetic progression, and for even n the superwords $\theta^n(a)$ start and end on a, so this can (and does) appear in a long monochromatic arithmetic progression.

By linear repetitivity of the Thue–Morse word v (see [13,15,16]), we know that $A(d)$ can be experimentally computed using a sufficiently long prefix of v. An algorithm can be defined to choose the length of this prefix in a finite number of steps. For a rigorous exposition the reader is referred to [1]. The data we have obtained following this idea confirms our results, including Theorems 15 and 21, and is presented in Fig. 1, which exhibits $A(d)$ for $d = 1, 2, \ldots, 1100$. A more comprehensive list of values of $A(d)$ can be found in [30].

The histogram in Fig. 2 counts, for each value y of $A(d)$, the number of d's that satisfy $A(d) = y$. We observe that short monochromatic arithmetic progressions are more frequent than long ones. If the range of differences d was larger, the peak of the histogram would be taller but roughly situated in the same place due to linear repetitivity of the Thue–Morse word. It seems that $A(d)$ never takes certain values and we conjecture that 3 is the smallest one among them.

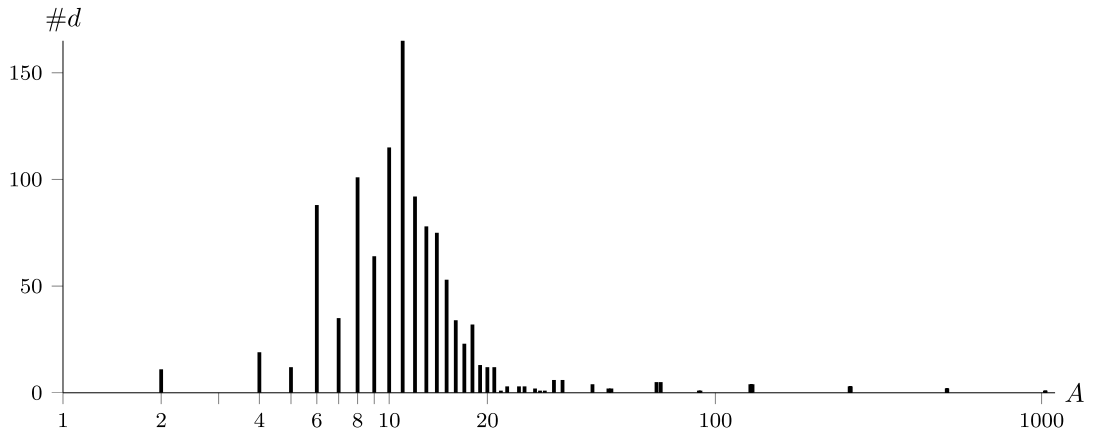


Fig. 2. Histogram corresponding to the values of Fig. 1

5. Generalised Thue–Morse words

Consider the generalised Thue–Morse substitution rules $\theta_{p,q}$ for $p, q \in \mathbb{N}^+$ defined by [9]

$$\theta_{p,q}: \begin{array}{l} 0 \mapsto 0^p 1^q \\ 1 \mapsto 1^p 0^q \end{array}, \tag{3}$$

where the original Thue–Morse substitution corresponds to $p = q = 1$. These binary bijective substitutions share many properties with the Thue–Morse substitution. In particular, we still have the 'bar-swap' symmetry $\theta^n(\overline{a}) = \overline{\theta^n(a)}$. This implies that, once again, superwords are uniquely determined as soon as you know a single of its letters. Note that, however, the symmetry of superwords is only preserved when $p = q$, with superwords for even n being symmetric while those for odd n being antisymmetric under reflection. The other main change is that, rather than working modulo 2, we now have to work modulo $Q := p + q$. Also, it is clear from the substitution rule (3) that the language of $\theta_{p,q}$ is $(Q+1)$-powerfree (in fact, $(Q+\varepsilon)$-powerfree for any $\varepsilon > 0$), generalising the cube-freeness (overlap-freeness) of the Thue–Morse case.

We note that Parshina also considered generalised Thue–Morse words [32], but in her work the generalisation is to larger alphabets. Here, we consider a generalisation of the Thue–Morse sequence along the lines of [9,10,22], restricting ourselves to the binary case.

Since the rule $\theta_{p,p}^2$ is symmetric under reflection, the corresponding language is reflection symmetric too. However, if $p \neq q$, reflection swaps the languages defined by $\theta_{p,q}$ and $\theta_{q,p}$. As we shall now show, each of these languages itself is not reflection symmetric.

Lemma 25. *For $p \neq q$, the languages $\mathcal{L}_{p,q}$ and $\mathcal{L}_{q,p}$ defined by the substitutions $\theta_{p,q}$ and $\theta_{q,p}$, respectively, are different, so $\mathcal{L}_{p,q} \neq \mathcal{L}_{q,p}$. In particular, for $a \in \{0,1\}$, the words $\overline{a}\, a^p\, \overline{a}^{q+1}$ belong to $\mathcal{L}_{p,q}$ but not to $\mathcal{L}_{q,p}$.*

Proof. It is easy to verify that $aa\overline{a} \in \mathcal{L}_{p,q}$ for any $p, q \in \mathbb{N}^+$. Now,

$$\theta_{p,q}(aa\overline{a}) = a^p\, \overline{a}^q\, a^p\, \overline{a}^q\, \overline{a}^p\, a^q = a^p\, \overline{a}^q\, a^p\, \overline{a}^{p+q}\, a^q,$$

so $\overline{a}\, a^p\, \overline{a}^{q+1} \in \mathcal{L}_{p,q}$ for all $p \neq q$. By reflection, $\overline{a}\, a^p\, \overline{a}^{q+1} \notin \mathcal{L}_{q,p}$ is equivalent to $\overline{a}^{q+1}\, a^p\, \overline{a} \notin \mathcal{L}_{p,q}$, which we are going to show now.

Noting that $p \neq q$ and that $\mathcal{L}_{p,q}$ can only contain strings of the type $\overline{a}a^m\overline{a}$ for $m \in \{p, q, p+q\}$, it follows by recognisability that $a^p\overline{a}$ in $\overline{a}^{q+1}a^p\overline{a}$ has to be the start of the level-1 superword $\theta_{p,q}(a)$. However, it then has to be preceded by the level-1 superword that ends in $\overline{a}$, which is again $\theta_{p,q}(a) = a^p\overline{a}^q$. This is clearly impossible, establishing the claim. □

Remark 26. Similarly, considering the word $a\overline{aa} \in \mathcal{L}_{p,q}$, with

$$\theta_{p,q}(a\overline{aa}) = a^p\, \overline{a}^q\, \overline{a}^p\, a^q\, \overline{a}^p\, a^q = a^p\, \overline{a}^{p+q}\, a^q\, \overline{a}^p\, a^q,$$

we can show that $\overline{a}^{p+1}\, a^q\, \overline{a} \in \mathcal{L}_{p,q}$, but is not a word in $\mathcal{L}_{q,p}$ for $p \neq q$.

Since all substitutions are binary bijective, it follows from [9] that they are not pure point diffractive, which implies that they do not have pure point dynamical spectrum either. Hence, by Proposition 8, they cannot contain infinitely long monochromatic arithmetic progressions for any finite difference d. This means that we can again define the maximum length of a monochromatic arithmetic progression.

Definition 27. For a positive integer d, let $A_{p,q}(d)$ denote the maximum length of a monochromatic arithmetic progression of difference d within the generalised Thue–Morse word, fixed point of the substitution of Equation (3).

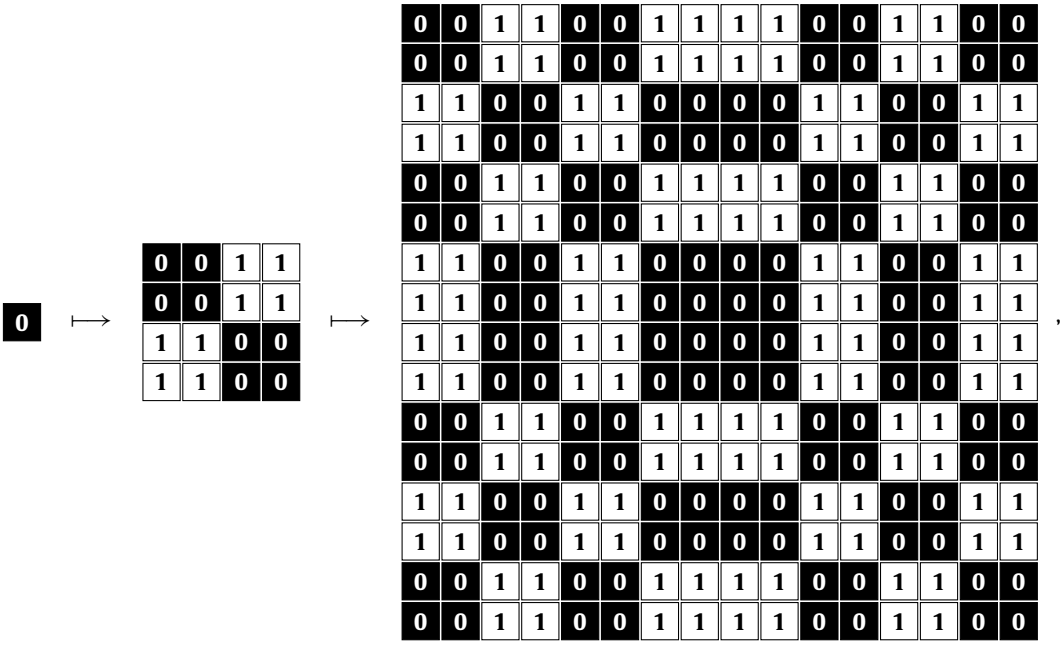
As a direct consequence of the substitution structure and recognisability, we know that $A_{p,q}(Q^n d) = A_{p,q}(d)$ holds for all $n \in \mathbb{N}$. In particular, this implies that $A_{p,q}(Q^n) = A_{p,q}(1) = Q$. We can again find long monochromatic arithmetic progressions by considering a block substitution.

Definition 28. Let $\Theta_{p,q}$ be the block substitution on the alphabet $\{0, 1\}$ defined by

$$\Theta_{p,q}\colon \quad 0 \longmapsto \begin{matrix} 0^p1^q \\ 0^p1^q \\ \vdots \\ 0^p1^q \\ 1^p0^q \\ 1^p0^q \\ \vdots \\ 1^p0^q \end{matrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\\a\end{matrix}}\right\} p \\ \\ \\ \left.\vphantom{\begin{matrix}a\\a\\a\\a\end{matrix}}\right\} q \end{matrix} , \quad 1 \longmapsto \begin{matrix} 1^p0^q \\ 1^p0^q \\ \vdots \\ 1^p0^q \\ 0^p1^q \\ 0^p1^q \\ \vdots \\ 0^p1^q \end{matrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\\a\end{matrix}}\right\} p \\ \\ \\ \left.\vphantom{\begin{matrix}a\\a\\a\\a\end{matrix}}\right\} q \end{matrix} .$$

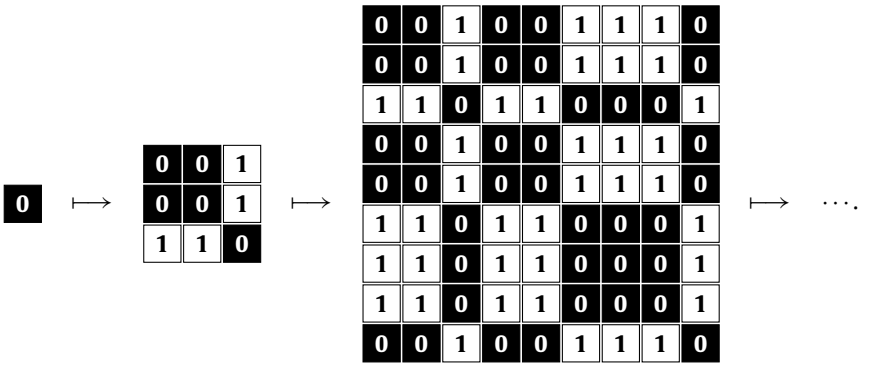
The block substitution $\Theta_{p,q}$ maps a single letter a to a $Q \times Q$ block of letters, which, when read line by line, coincides with the word $\theta^2_{p,q}(a)$. To illustrate the properties of $\Theta_{p,q}$, let us consider a couple of examples.

Example 29. We first consider and example where $p = q$, namely $\Theta_{2,2}$. The first two substitution steps of the letter 0 are as follows,



which is a similar structure as for the original Thue–Morse case. In particular, all squares along the diagonals are of the same colour.

The situation is different for $p \neq q$. Here, we consider the block substitution $\Theta_{2,1}$ as an example, which acts on a letter 0 as



Note that, while we retain the same letter along the main diagonal, this is no longer the case along the diagonal from the lower left to the top right. Considering the second inflation step shown above, it appears that there is a long monochromatic progression of difference $3^2-1=8$, starting from the central black square on the top row and moving down diagonally, and then continuing on from the final black square on the middle row. Indeed, we find that for $d=8$ the longest monochromatic arithmetic progression has length 12; however, this pattern does not persist for further inflation steps.

As for the Thue–Morse case, the image of a letter a has all entries a along the main diagonal of this block. However, as illustrated in Example 29, in general this is no longer the case for the other diagonal, except for the case that $p=q$, in which case the entries of this diagonal are all $\overline{a}$. This means that we obtain the existence of long monochromatic arithmetic progressions, as in the Thue–Morse case, for $d=Q^n+1$ for all values of p and q, while long monochromatic arithmetic progressions for $d=Q^n-1$ may only exist if $p=q$.

Noting that Lemma 17 and Lemma 18 generalise in a straightforward manner, we obtain the following existence result for long monochromatic arithmetic progressions in generalised Thue–Morse words, generalising the result of Lemma 19.

Lemma 30. *For all $n, p, q \in \mathbb{N}^+$, $Q=p+q$, we have that $A_{p,q}(Q^n+1) \geqslant Q^n$ and $A_{p,p}(Q^n-1) \geqslant Q^n$. If n is even, we further have that $A_{p,p}(Q^n-1) \geqslant Q^n+2$.*

Proof. This follows by the same line of argument as for the Thue–Morse case in the proof of Lemma 19. □

The following results implicitly use the fact that, as in the Thue–Morse case, a level-n superword of $\theta_{p,q}$ within any word in $\mathcal{L}_{p,q}$ only occurs in certain ways. We obtain the following generalisation of Lemma 20.

Lemma 31. *The word $w=\theta_{p,q}^n(a)$ with $a \in \{0,1\}$ and $n>1$ occurs inside sufficiently long words in $\mathcal{L}_{p,q}$ either as the level-n superword itself, or, in the case when $p=q$, in the centre of two level-n superwords $\theta_{p,p}^n(\overline{a})\theta_{p,p}^n(\overline{a})$. In the latter case $p=q$, if w is followed by the letter $\overline{a}$ or preceded by the letter a (for n odd) or $\overline{a}$ (for n even), it is the level-n superword.*

Proof. The proof is a straightforward generalisation from that of Lemma 20. The only difference is that, for $p \neq q$, the word w can only occur as the level-n superword, because the level-$(n-1)$ superwords are determined and with $p \neq q$ they can only be combined to the level-n superword in one way. □

Theorem 32. *For all $n, p, q \in \mathbb{N}^+$, $Q=p+q$, $n>1$, we have that*

$$A_{p,q}(Q^n+1)=\begin{cases} Q^n+Q-2, & \text{if } p>1 \text{ and } q>1,\\ Q^n+Q-1, & \text{if } q>p=1 \text{ or } p>q=1,\\ Q^n+Q, & \text{if } p=q=1. \end{cases}$$

Proof. From the proof of Lemma 30, we have an arithmetic progression of length Q^n of the letter a in the word

$$w=\theta_{p,q}^{2n}(a)=\left(\theta_{p,q}^n(a)\right)^p\left(\theta_{p,q}^n(\overline{a})\right)^q\ldots\left(\theta_{p,q}^n(\overline{a})\right)^p\left(\theta_{p,q}^n(a)\right)^q,$$

with the first and final letter being part of the progression (as before, because we are looking at an even number of substitutions, the word w starts and ends with the same letter). Note that, since the elements in the progression of difference Q^n+1 visit successive positions in superwords $\theta_{p,q}^n(a)$ in order, we know that, irrespective of where we start, once we hit the first letter of a superword $\theta_{p,q}^n(a)$ (which has to happen for any progression of length Q^n) the progression follows this same sequence, and the same backwards from when we hit the final position in a level-n superword. Using the same argument as in the proof of Theorem 21, we conclude that any progression of length $L>Q^n$ has to include this superword.

Now consider how many letters can be added at either end of the progression of length Q^n in the superword w. For $Q > 2$, all four possibilities for this superword being bordered by level-n superwords $u = \theta^n_{p,q}(a)$ or $\overline{u} = \theta^n_{p,q}(\overline{a})$ can occur, so we need to consider w followed or preceded by either u or $\overline{u}$.

If w is followed by u, it is followed by u^p and we can extend the arithmetic progression by exactly $p-1$ to the right. If it is followed by $\overline{u}$, we cannot extend at all unless $p = 1$. For $p = 1$, we can extend by exactly one step.

If w is preceded by u (for n odd) or $\overline{u}$ (for n even), we cannot extend at all unless $q = 1$, in which case we can extend by precisely one step. If it is preceded by $\overline{u}$ (for n odd) or u (for n even), we can extend by exactly $q-1$ steps to the left.

Choosing the combination with the longest available progression yields the result. □

Note that for $p = q = 1$ we recover the result of Theorem 21.

Theorem 33. *For all n, $p \in \mathbb{N}^+$, $Q = 2p$, $n > 1$, we have that*

$$A_{p,p}(Q^n-1) = \begin{cases} Q^n, & \text{if } n \text{ is odd,} \\ Q^n + Q, & \text{if } n \text{ is even and } p > 1, \\ Q^n + Q + 2, & \text{if } n \text{ is even and } p = 1. \end{cases}$$

Proof. From Lemma 30, we already know that long monochromatic arithmetic progressions for $d = Q^n - 1$ exist, with $A_{p,p}(d) \geqslant Q^n$, within the superword

$$w = \theta^{2n}_{p,p}(a) = \left(\theta^n_{p,p}(a)\right)^p \left(\theta^n_{p,p}(\overline{a})\right)^p \ldots \left(\theta^n_{p,p}(\overline{a})\right)^p \left(\theta^n_{p,p}(a)\right)^p.$$

Accordingly, such a long progression visits every position in level-n superwords.

For odd values of n, the superwords $\theta^n_{p,p}(b)$ start with b and end on $\overline{b}$, so it is not possible to have the first and last letter in the same monochromatic arithmetic progression. This implies that $A_{p,p}(d) \leqslant Q^n$, and hence $A_{p,p}(d) = Q^n$ in this case.

For even values of n, all superwords $\theta^n_{p,p}(b)$ start and end in the same letter, and hence we have $A_{p,p}(d) \geqslant Q^n + 2$ as shown in Lemma 30, with the first and last letter in the superword w belonging to the arithmetic progression. What is left to consider is how far this can be extended on either side. The word w can be preceded and succeeded by level-n superwords $u = \theta^n_{p,p}(a)$ or $\overline{u} = \theta^n_{p,p}(\overline{a})$, where for $p = 1$ one has to ensure cube-freeness.

If w is succeeded by u and hence by u^p, it can be extended by exactly $p-1$ steps. If it is succeeded by $\overline{u}$, no extension is possible, unless $p = 1$ in which case you can extend by exactly one step. Due to symmetry of all these words for even n, the same argument applies at the other end, which completes the proof. □

Proposition 34. *For all n, $p, q \in \mathbb{N}^+$, with $n > 2$, $p \neq q$ and $Q = p + q$, we have that $A_{p,q}(Q^n-1) \leqslant Q^n$.*

Proof. Assume to the contrary that a long monochromatic arithmetic progression of difference $Q^n - 1$ and length $L > Q^n$ exists. Then this progression contains a level-n superword $w = \theta^n_{p,q}(a)$ with two instances of this progression, implying that the first and last letter of w agree. If n is odd, this is not possible, since w starts with a and ends on $\overline{a}$.

If $n > 2$ is even, w starts and ends with

$$\theta^2(a) = \theta(a)^p \theta(\overline{a})^q = (a^p \overline{a}^q)^p (\overline{a}^p a^q)^q.$$

Since by bijectivity a single letter determines the superwords, we can read off the sequence of words to the left and to the right of the word w with two instances of the progression, provided the progression extends.

Consider first the case $p > 1$. Assume that the progression continues to the right of w. As we are considering the difference $d = Q^n - 1$, we are effectively reading the word w "backwards" to determine the sequence of superwords that is required. As mentioned above, w ends on $\theta^2(a)$ which (since $p > 1$) contains the word $\overline{a}\, a^p\, \overline{a}^{p+q}$. According to Lemma 25, this word does not occur in $\mathcal{L}_{p,q}$, since we are considering the case that $p \neq q$. This implies that the sequence of superwords required to continue the progression for Q^2 steps to the right contains a subsequence that corresponds to the images of a word under $\theta_{p,q}$ that is not in the language $\mathcal{L}_{p,q}$, which is a contradiction. This means that the progression cannot continue to the right for more than $(q+1)Q$ steps at most.

An analogous argument holds if you assume that the progression extends to the left, showing that it can at most continue for pQ steps to the left. So the total length of the progression is at most $Q^2 + Q < Q^n - 1$ for $n > 2$.

If $p = 1$ and hence $q > 1$, we can use the same arguments as above, based on the word $\overline{a}^{p+1}\, a^q\, \overline{a}$ from Remark 26, which occurs within $\theta^2_{p,q}(a)$ in this case. □

So we have established the existence of long monochromatic arithmetic progressions for all generalised Thue–Morse sequences for differences $d = Q^n + 1$, as well as for differences $d = Q^n - 1$ in the case that $p = q$. The obvious conjecture is that these are again the longest monochromatic arithmetic progressions that you can find, up to the given difference, in these systems, which we state as a conjecture.

Conjecture 35. *For all n, $p, q \in \mathbb{N}^+$, $Q = p + q$, $n > 2$, we have that*

$$\max_{d \leqslant Q^n+1} A_{p,q}(d) = \begin{cases} A_{p,q}(Q^n-1) = Q^n + Q + 2, & \text{if } p = q = 1 \text{ and } n \text{ even,} \\ A_{p,q}(Q^n-1) = Q^n + Q, & \text{if } p = q > 1 \text{ and } n \text{ even,} \\ A_{p,q}(Q^n+1) = Q^n + Q, & \text{if } p = q = 1 \text{ and } n \text{ odd,} \\ A_{p,q}(Q^n+1) = Q^n + Q - 1, & \text{if } q > p = 1 \text{ or } p > q = 1, \\ A_{p,q}(Q^n+1) = Q^n + Q - 2, & \text{if } p, q > 1, \text{ and } p \neq q \text{ or } n \text{ odd.} \end{cases}$$

To establish this conjecture, we would need to generalise the results of Lemmas 23 and 24. This is not straightforward, though, because we now have to consider differences $d = Q^n - k$ where we may have that k is a non-trivial divisor of Q, in which case the argument that in a long arithmetic progression all rest classes modulo Q^n appear is no longer applicable.

The following lemma details the relations within superwords arising from an assumed existence of long monochromatic arithmetic progressions.

Lemma 36. *Consider the language of the generalised Thue–Morse substitution $\theta_{p,q}$ with $p + q = Q$. For $n \in \mathbb{N}$, $n > 1$, and $1 < k < Q^{n-1}$, set $s = \gcd(k, Q)$ and assume that $s \neq Q$. If there exists a long monochromatic arithmetic progression of difference $d = Q^n - k$ and length $L > Q^n/s$, the level-n superwords $w = \theta^n_{p,q}(a)$ have to satisfy $w_{r+\ell s} = w_{r+\ell s+d}$ for some $0 \leqslant r < s$ and for all $0 \leqslant \ell < k/s$.*

Proof. We have $\gcd(d, Q) = \gcd(k, Q) = s$ and $r \equiv d \bmod Q$, so any such arithmetic progression of length L visits all positions

$$\left\{r' \mid 0 \leqslant r' < Q^n, r \equiv r' \bmod Q^n\right\} = \left\{r + \ell s \mid 0 \leqslant \ell < \tfrac{Q^n}{s}\right\}$$

in a superword $w = \theta^n_{p,q}(a)$ for some letter $a \in \{0, 1\}$ and for some $0 \leqslant r < s$. Whenever there are two instances within a superword, the corresponding letters have to agree for either superword, since $\theta^n_{p,q}(\overline{a}) = \overline{\theta^n_{p,q}(a)}$. The condition for having two instances within a superword is $r' + d < Q^n$, which means $r' < k$. With $r' = r + \ell s$, this results in $\ell < (k - r)/s$, and since $r < s$ this is equivalent to $\ell < k/s$, establishing the claim. □

Proposition 37. *Consider the language of the generalised Thue–Morse substitution $\theta_{p,q}$ for $Q = p + q$ prime. Then, for $n \in \mathbb{N}$, $n > 1$, and $Q < k < Q^{n-1}$, any monochromatic arithmetic progression of difference $d = Q^n - k$ has length $L \leqslant Q^n$.*

Proof. Since Q is prime, we have that $\gcd(k, Q) = 1$. From Lemma 36 we know that, if there exists a monochromatic arithmetic progression of length $L > Q^n$, the superwords $w = \theta^n_{p,q}(a)$ have to satisfy $w_{[0,k)} = w_{[Q^n-k, Q^n)}$. If $k > Q$, this produces a contradiction, since the final Q letters of $w_{[0,k)}$ cannot be a valid level-1 superword, but w has to end on a level-1 superword. □

Note that this lemma does not cover the differences $d = Q^n - k$ where $1 < k < Q$, which we would need to establish the conjecture for prime values of Q (except for $Q = 2$ which brings us back to the Thue–Morse case). A partial result for $\min(p, q) = 1$ is next, establishing Conjecture 35 for this class.

Proposition 38. *Consider the language of the generalised Thue–Morse substitution $\theta_{p,q}$ for $\min(p, q) = 1$ and $Q = p + q$ prime. Then, for $n \in \mathbb{N}^+$ and any $1 < k < Q^{n-1}$, any monochromatic arithmetic progression of difference $d = Q^n - k$ has length $L \leqslant Q^n$.*

Proof. From Proposition 37, we know that the claim holds from $k > Q$.

If $p = 1$, the level-1 superwords are of the form $a\overline{a}^q$ with $a \in \{0, 1\}$. This means that, for $1 < k < Q$, the superwords $w = \theta^n_{1,q}(a)$ start with $w_{[0,k)} = a\overline{a}^{k-1}$. Since $1 \leqslant k - 1 < q$, this string of letters cannot occur at the end of the superword w.

Similarly, if $q = 1$, the superwords $w = \theta^n_{p,1}(a)$ start with $w_{[0,k)} = a^k$. Since $k > q = 1$, this string of letters cannot occur at the end of the superword w. □

We finish this section with plots of $A_{p,q}(d)$. To depict the difference between the $p = q$ case and the $p \neq q$ case, we present Figs. 3 and 4, the former for $p = q = 2$ and the latter for $p = 3$ and $q = 1$, so $Q = 4$ in both cases. The experimental data agree with our results, including Theorems 32 and 33, and also give credibility to Conjecture 35.

In Fig. 3, the equality of p and q preserves the symmetry of superwords of the Thue–Morse substitution and consequently, it is qualitatively similar to Fig. 1. In particular, we observe the large monochromatic arithmetic progressions at differences $d = Q^n \pm 1 = 4^n \pm 1$. It is interesting to note that there is another series of large peaks around differences d of the form 2^n for odd n, similarly to Fig. 1 for the Thue–Morse case; however, the largest values are not at $d = 2^n \pm 1$, but at $d = 2^n - 2$. On the other hand, when $p \neq q$, the output differs qualitatively from the Thue–Morse case, as it can be seen in Fig. 4.

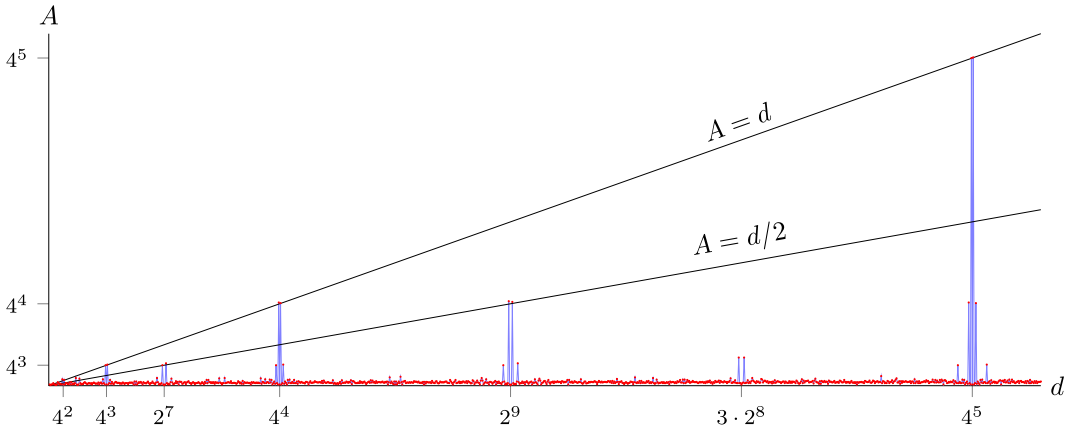


Fig. 3. $A_{2,2}(d)$ for $d = 1, 2, \ldots, 1100$.

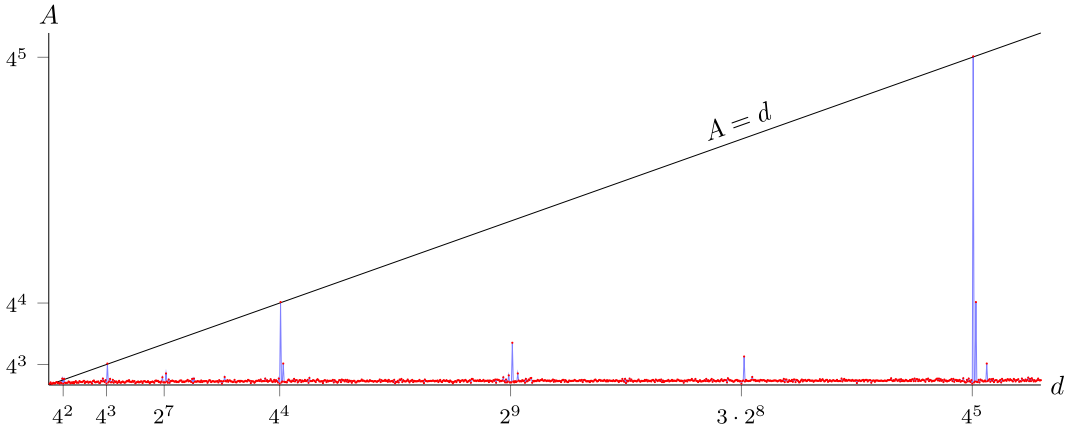


Fig. 4. $A_{3,1}(d)$ for $d = 1, 2, \ldots, 1100$.

6. Some upper bounds for $A(d)$ for primitive binary bijective substitutions

In this section, we state a partial result for general primitive bijective (and hence constant-length) substitutions over a binary alphabet $\mathcal{A}$, which include those treated in Section 5.

For a finite or infinite word $v = (v_i)$, a finite word $w = w_0 w_1 \cdots w_n$ and a natural number k, we say that a pattern $\mathrm{P}(w, k)$ is legal in v if there is an s such that

$$v_s = w_0,\ v_{s+k} = w_1,\ v_{s+2k} = w_2,\ \ldots,\ v_{s+nk} = w_n.$$

Namely, $\mathrm{P}(w, k)$ is an arithmetic subword (not necessarily monochromatic) in v of difference k and length $n+1$. A monochromatic progression corresponds to the case when $w = 0^{n+1}$ or 1^{n+1}. Note that, if we say a $\mathrm{P}(w, k)$ is legal in $v = (v_i)$, we assume that $s, s+k, \ldots, s+nk$ are in the range of the subscripts i for v. In this section, we let ϱ be a primitive binary bijective substitution rule of length Q on an alphabet $\mathcal{A} = \{0, 1\}$ and $v = (v_n)_{n\in\mathbb{N}}$ be a fixed point of ϱ.

Proposition 39. *Let $w = w_0 w_1 \cdots w_n$ be a finite word on $\mathcal{A}$ and k, d, m be positive integers. Assume that*

1. *Q and d are coprime,*
2. *$\mathrm{P}(w, k)$ is not legal in v, and*
3. *$\mathrm{P}(w, d)$ is legal in $\varrho^m(0)$.*

Then $\mathrm{P}(0^{n'}, d')$ is not legal in v, where $n' = Q^m + n$ and $d' = Q^m k + d$. In other words, there are no monochromatic arithmetic progressions of difference d' and length n'.

Proof. Assume that $\mathrm{P}(0^{n'}, d')$ is legal in v. Then there is a $t \in \mathbb{N}$ such that

$$v_{sd'+t} = 0,$$

for each $s = 0, 1, \ldots n' - 1$.

By assumption 3 in the statement, there is an $l_0 \in \mathbb{N}$ such that the $(jd+l_0)$th letter in $\varrho^m(0)$ is w_j for any $j = 0, 1, \ldots, n$. Note that this implies that we have $0 \leqslant jd + l_0 \leqslant Q^m - 1$ for each $j = 0, 1, \ldots, n$.

Since Q and d are coprime (assumption 1 in the statement), there is an $s \in \{0, 1, \ldots Q^m - 1\}$ such that $l_0 \equiv t + sd' \bmod Q^m$, and so there is an $i \in \mathbb{N}$ such that $l_0 + iQ^m = t + sd'$. For each $j = 0, 1, \ldots, n$, we have that

$$t + (s + j)d' = (i + kj)Q^m + jd + l_0.$$

Since $0 \leqslant jd + l_0 \leqslant Q^m - 1$, $v_{t+(s+j)d'}$ is the $(jd + l_0)$th letter in a superword, which is denoted by $\varrho^m(a_j)$.

If $w_j = 0$, then the $(jd+l_0)$th letter in $\varrho^m(0)$ is 0 and so the $(jd+l_0)$th letters in $\varrho^m(0)$ and $\varrho^m(a_j)$ coincide. This means that $a_j = 0$.

If $w_j = 1$, then the $(jd + l_0)$th letter in $\varrho^m(0)$ is 1 and so the $(jd + l_0)$th letters in $\varrho^m(0)$ and $\varrho^m(a_j)$ differ. This means that $a_j = 1$.

In any cases, $w_j = a_j$ and we have that

$$v_{i+kj} = w_j$$

for each $j = 0, 1, \ldots, n$, and $\mathrm{P}(w, k)$ appears in v. However, this contradicts assumption 2 in the statement. □

In the next proposition, we use the notation $\tilde{w}$ for the palindromic inverse $w_n w_{n-1} \cdots w_1$ of a finite word $w = w_0 w_1 \cdots w_n$.

Proposition 40. *Let $w = w_0 w_1 \cdots w_n$ be a finite word on $\mathcal{A}$ and k, d, m be positive integers. Assume that*

1. *Q and d are coprime,*
2. *$\mathrm{P}(w, k)$ is not legal in v, and*
3. *$\mathrm{P}(\tilde{w}, d)$ is legal in $\varrho^m(0)$.*

Then $\mathrm{P}(0^{n'}, d')$ is not legal in v, where $n' = Q^m + n$ and $d' = Q^m k - d$. In other words, there are no monochromatic arithmetic progressions of difference d' and length n'. □

We omit the proof, which is a modification of the proof of Proposition 39. This could be done by replacing $jd + l_0$ with $(n - j)d + l_0$ in the second paragraph and l_0 with $l_0 + nd$ in the subsequent paragraphs.

Example 41. Let ϱ be such that

$$\varrho: \begin{array}{l} 0 \mapsto 0101 \\ 1 \mapsto 1010. \end{array}$$

Set $w = 101$. Then for each $m = 1, 2, \ldots$, w appears in $\varrho^m(0)$ and so $\mathrm{P}(w, 1)$ is legal in $\varrho^m(0)$. On the other hand, $\mathrm{P}(w, 2)$ is not legal in the fixed point v. Therefore, by Propositions 39 and 40 with $Q = 4$, $d = 1$ and $k = 2$, the monochromatic arithmetic progressions of difference $d' = 2Q^m \pm 1$ and length $n' = Q^m + 2$ are not legal in v.

7. Conclusions and outlook

We investigated the occurrence of long monochromatic arithmetic progressions in the Thue–Morse word and a class of generalised Thue–Morse words. Clearly, the existence of these long progressions of difference d and length $L \approx d$ was linked to the structure of the underlying substitution, corresponding to a diagonal in the induced block substitution carrying the same letter. If we change the order of letters in the substitution while retaining bijectivity, this property will be lost and, in general, such long progressions should not be expected to exist. We currently do not have any stronger bounds on the progressions for these cases, but it is likely that infinite series of progressions where the length L grows linearly with the difference d may not exist in this case. It would be interesting to quantify the behaviour for such systems. In this vein, we recall Conjecture 35 about the maximal values of monochromatic arithmetic progressions in generalised binary Thue–Morse words.

The behaviour of the maximum length of monochromatic arithmetic progressions is rather volatile, as can be seen from the plots for $A(d)$ in this paper. It is not even known whether $A(d)$ can take all values; we believe not, and pose the following:

Conjecture 42. *There is no difference d_0 for which the maximum length of a monochromatic arithmetic progression $A(d_0)$ is 3 in the Thue-Morse sequence.*

A complete justification of the theory behind our computations can be found in [1].

Other potential generalisations include the investigation of other substitution sequences, either with more letters (see [32] for results on a different class of generalised Thue–Morse sequences) or for non-bijective substitutions, or quantitative versions of normality results [7,28,34,35].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Neil Mañibo for fruitful discussions and valuable suggestions. We also thank the referees for helpful comments.

References

[1] I. Aedo, PhD thesis: Forward limit sets of semigroups of substitutions and monochromatic arithmetic progressions in automatic sequences, The Open University, (UK), in preparation.
[2] I. Aedo, U. Grimm, N. Mañibo, Y. Nagai, P. Staynova, Monochromatic arithmetic progressions in automatic sequences, in preparation.
[3] J.-P. Allouche, J. Shallit, The ubiquitous Prouhet–Thue–Morse sequence, in: C. Ding, T. Helleseth, H. Niederreiter (Eds.), Sequences and Their Applications, Springer, London, 1999, pp. 1–16.
[4] J.-P. Allouche, J. Shallit, Sums of digits, overlaps, and palindromes, Discret. Math. Theor. Comput. Sci. 4 (1) (2000) 1–10.
[5] J.-P. Allouche, J. Shallit, Automatic Sequences: Theory, Applications, Generalizations, Cambridge University Press, Cambridge, 2003.
[6] S.V. Avgustinovich, J. Cassaigne, A.E. Frid, Sequences of low arithmetical complexity, RAIRO Theor. Inform. Appl. 40 (2006) 569–582.
[7] S.V. Avgustinovich, D.G. Fon-Der-Flaass, A.E. Frid, Arithmetical complexity of infinite words, in: M. Ito, T. Imaoka (Eds.), Words, Languages & Combinatorics III, World Scientific, Singapore, 2003, pp. 51–62.
[8] M. Baake, F. Gähler, Pair correlations of aperiodic inflation rules via renormalisation: some interesting examples, Topol. Appl. 205 (2016) 4–27.
[9] M. Baake, F. Gähler, U. Grimm, Spectral and topological properties of a family of generalised Thue–Morse sequences, J. Math. Phys. 53 (2012) 032701.
[10] M. Baake, U. Grimm, Surprises in aperiodic diffraction, J. Phys. Conf. Ser. 226 (2010) 012023.
[11] M. Baake, U. Grimm, Aperiodic Order. Vol. 1: A Mathematical Invitation, Cambridge University Press, Cambridge, 2013.
[12] J. Cassaigne, A.E. Frid, On the arithmetical complexity of Sturmian words, Theor. Comput. Sci. 380 (2007) 304–316.
[13] D. Damanik, D. Lenz, Substitution dynamical systems: characterization of linear repetitivity and applications, J. Math. Anal. Appl. 321 (2006) 766–780.
[14] F.M. Dekking, The spectrum of dynamical systems arising from substitutions of constant length, Z. Wahrscheinlichkeitstheor. Verw. Geb. 41 (1978) 221–239.
[15] F. Durand, A characterization of substitutive sequences using return words, Discrete Math. 179 (1998) 89–101.
[16] F. Durand, B. Host, C. Skau, Substitutive dynamical systems, Bratteli diagrams and dimension groups, Ergod. Theory Dyn. Syst. 19 (1999) 953–993.
[17] A.E. Frid, Arithmetical complexity of symmetric D0L words, Theor. Comput. Sci. 306 (2003) 535–542.
[18] A.E. Frid, Sequences of linear arithmetic complexity, Theor. Comput. Sci. 339 (2005) 68–87.
[19] A.E. Frid, Prefix palindromic length of the Thue–Morse word, Preprint, arXiv:1906.09392, 2019.
[20] A.O. Gelfond, Sur les nombres qui ont des propiétés additives et multiplicatives données, Acta Arith. 13 (1968) 259–265.
[21] K. Jacobs, M. Keane, $0-1$ sequences of Toeplitz type, Z. Wahrscheinlichkeitstheor. Verw. Geb. 13 (1969) 123–131.
[22] M. Keane, Generalized Morse sequences, Z. Wahrscheinlichkeitstheor. Verw. Geb. 10 (1968) 335–353.
[23] J. Konieczny, Gowers norms for the Thue–Morse and Rudin–Shapiro sequences, Ann. Inst. Fourier 69 (2019) 1897–1913.
[24] M. Lothaire, Algebraic Combinatorics on Words, Cambridge University Press, 2002.
[25] A. de Luca, E.V. Pribavkina, L.Q. Zamboni, A coloring problem for infinite words, J. Comb. Theory, Ser. A 125 (2014) 306–332.
[26] J.C. Martin, Substitution minimal flows, Am. J. Math. 93 (1971) 503–526.
[27] J.F. Morgenbesser, J. Shallit, T. Stoll, Thue–Morse at multiples of an integer, J. Number Theory 131 (2011) 1498–1512.
[28] C. Müllner, L. Spiegelhofer, Normality of the Thue–Morse sequence along Piatetski–Shapiro sequences, II, Isr. J. Math. 220 (2017) 691–738.
[29] Y. Nagai, S. Akiyama, J-Y. Lee, On arithmetic progressions in non-periodic self-affine tilings, Ergod. Theory Dyn. Syst. (2021) 1–33.
[30] O.E.I.S. Foundation Inc, Entry A342818 in the on-line encyclopedia of integer sequences, https://oeis.org/A342818, 2021.
[31] O.G. Parshina, On arithmetic progressions in the generalized Thue–Morse word, in: F. Manea, D. Nowotka (Eds.), WORDS 2015, in: Lecture Notes in Computer Science, vol. 9304, Springer, Cham, 2015, pp. 191–196.
[32] O.G. Parshina, On arithmetic index in the generalized Thue–Morse word, in: S. Brlek, F. Dolce, C. Reutenauer, É. Vandomme (Eds.), WORDS 2017, in: Lecture Notes in Computer Science, vol. 10432, Springer, Cham, 2017, pp. 121–131.
[33] O.G. Parshina, ПЕРИОДИЧЕСКИЕ СТРУКТУРЫ В МОРФИЧЕСКИХ СЛОВАХ И РАСКРАСКАХ БЕСКОНЕЧНЫХ ЦИРКУЛЯНТНЫХ ГРАФОВ, PhD thesis, Université de Lyon and Sobolev Institute of Mathematics, 2019.
[34] L. Spiegelhofer, Normality of the Thue–Morse sequence along Piatetski–Shapiro sequences, Q. J. Math. 66 (2015) 1127–1138.
[35] L. Spiegelhofer, The level of distribution of the Thue–Morse sequence, Compos. Math. 156 (12) (2020) 2560–2587.
[36] B.L. van der Waerden, Beweis einer Baudetschen Vermutung, Nieuw Arch. Wiskd. 15 (1927) 212–216.
[37] C. Wojcik, L.Q. Zamboni, Colouring problems for infinite words, in: V. Berthé, M. Rigo (Eds.), Sequences, Groups, and Number Theory, Birkhäuser, Basel, 2018, pp. 213–231.