J. Aust. Math. Soc. **114** (2023), 110–144
doi:10.1017/S1446788721000380

GAPS IN THE THUE–MORSE WORD

LUKAS SPIEGELHOFER

(Received 8 February 2021; accepted 14 September 2021; first published online 25 January 2022)

Communicated by Michael Coons

Abstract

The Thue–Morse sequence is a prototypical automatic sequence found in diverse areas of mathematics, and in computer science. We study occurrences of factors w within this sequence, or more precisely, the sequence of gaps between consecutive occurrences. This gap sequence is morphic; we prove that it is not automatic as soon as the length of w is at least 2, thereby answering a question by J. Shallit in the affirmative. We give an explicit method to compute the *discrepancy* of the number of occurrences of the block $\mathtt{01}$ in the Thue–Morse sequence. We prove that the sequence of discrepancies is the sequence of output sums of a certain base-2 transducer.

2020 *Mathematics subject classification*: primary 68Q45, 68R15; secondary 11A63.

Keywords and phrases: Thue–Morse sequence, subword occurrence, morphic sequence, k-kernel.

1. Introduction and main result

Automatic sequences can be defined via deterministic finite automata with output: feeding the base-q expansion (where $q \geq 2$ is an integer) of $0, 1, 2, \ldots$ into such an automaton, we obtain an automatic sequence as its output, and each automatic sequence is obtained in this way. One of the simplest automatic sequences (in terms of the size of the defining substitution) is the Thue–Morse sequence $\mathbf{t}$. It is the fixed point of the substitution τ given by

$$\tau : \mathtt{0} \mapsto \mathtt{01}, \quad \mathtt{1} \mapsto \mathtt{10}, \tag{1-1}$$

starting with $\mathtt{0}$:

$$\mathbf{t} = \tau^{\omega}(\mathtt{0}) = \mathtt{0110100110010110100101100110100}1 \cdots . \tag{1-2}$$

(Here $\tau^{\omega}(\mathtt{0})$ denotes the pointwise limit of the iterations $\tau^k(\mathtt{0})$, in symbols $\tau^{\omega}(\mathtt{0})|_j = \lim_{k\to\infty} \tau^k(\mathtt{0})|_j$. We use analogous notation in other places too.) Occurrences of

The author was supported by the Austrian Science Fund (FWF), project F5502-N26, which is a part of the Special Research Program 'Quasi Monte Carlo methods: Theory and Applications', and by the FWF-ANR project ArithRand, grant numbers I4945-N and ANR-20-CE91-0006.

this sequence in different areas of mathematics can be found in the paper [3] by Allouche and Shallit, which also offers a good bibliography. Another survey paper on the Thue–Morse sequence was written by Mauduit [23]. Much more information concerning automatic and general morphic sequences is presented in the book [4] by Allouche and Shallit.

Terms 1–12 of the Thue–Morse sequence (the 0 th term is omitted) make a peculiar appearance in *The Simpsons Movie* [12]:

> Russ Cargill: I want 10,000 tough guys, and I want 10,000 soft guys to make the tough guys look tougher! And here's how I want them arranged: tough, tough, soft, tough, soft, soft, tough, tough, soft, soft, tough, soft.

This could, of course, be a coincidence. A different, more sensible explanation of this appearance is along the lines of Brams and Taylor [11, pages 36–44]. They rediscover the Thue–Morse sequence while seeking *balanced alternation* between two parties, 'Ann' and 'Ben'. However, Brams and Taylor do not attribute the resulting sequence to Thue and Morse (nor to Prouhet).

We are interested in counting the number of times that a word occurs as a *factor* (a contiguous finite subsequence) of another word; a related concept is the *binomial coefficient of two words*, which counts the corresponding number concerning general subsequences. We wish to note the related paper by Rigo and Salimov [25], defining the *m-binomial equivalence* of two words, and the later paper by Lejeune *et al.* [21], where the Thue–Morse sequence is investigated with regard to this new concept.

It is well known that the sequence **t** is *uniformly recurrent* [22, Section 1.5.2]. That is, for each factor w of **t** there is a length n with the following property: every contiguous subsequence of **t** of length n contains w as a factor. The factor `01` therefore appears in **t** with bounded distances. We are interested in the infinite word **B** (over a finite alphabet) describing these differences. We mark the occurrences of `01` in the first 64 letters of **t**:

$$\mathtt{011010011001011010010110011010011001011}\,y\,\mathtt{011010010110100110010110}, \tag{1-3}$$

from which we see that

$$\mathbf{B} = \mathtt{33423324334243323342 3}\cdots.$$

The blocks `000` and `111` do not appear as factors in **t**, since **t** is a concatenation of the blocks `01` and `10`. Therefore the gaps between consecutive occurrences of `01` in **t** are in fact bounded by 4, and clearly they are bounded below by 2 (since different occurrences of `01` cannot overlap). It follows that we only need three letters, `2`, `3`, and `4`, in order to capture the gap sequence.

The set of *return words* [16, 20] of a factor w of **t** is the set of words x of the form $x = w\tilde{x}$, where w is not a factor of $\tilde{x}$, and $w\tilde{x}w$ is a factor of **t**. The gap sequence is the

sequence of lengths of words in the decomposition $\mathbf{t} = x_0 x_1 \cdots$ of the Thue–Morse sequence into return words of $\mathtt{01}$, which are $\mathtt{011}$, $\mathtt{010}$, $\mathtt{0110}$, and $\mathtt{01}$ in order of appearance.

An appearance of the factor $\mathtt{01}$ marks the beginning of a block of 1s in $\mathbf{t}$. Moreover, no other block of 1s can appear before the next appearance of $\mathtt{01}$: between two blocks of 1s we can find a block of one or more $\mathtt{0}$s, and the last $\mathtt{0}$ in this block is followed by 1. The assumption that we see a block of 1s before the next appearance of $\mathtt{01}$ therefore leads to a contradiction. This argument is clearly visible in (1-3). The sequence $\mathbf{B}$ therefore gives the distances of consecutive blocks of 1s. We see in Lemma 2.3 that the sequence $\mathbf{B}$ is *morphic* or *substitutive*. That is, it can be described as the coding of a fixed point of a substitution over a finite alphabet. Jeffrey Shallit (private communication, July 2019) proposed to prove the nonautomaticity of $\mathbf{B}$ to the author. In the present paper, we investigate the sequence $\mathbf{B}$ and the closely related, very well-known automatic sequence $\mathbf{A}$ defined in Section 2.1. In particular, we prove the following theorem.

THEOREM 1.1. *Let w be a factor of the Thue–Morse word of length at least 2, and C the sequence of gaps between consecutive occurrences of w in $\mathbf{t}$. Then C is morphic, but not automatic.*

Note that the set of positions where a given factor w appears in $\mathbf{t}$ *is* 2-automatic; that is, its characteristic sequence is automatic. This follows from the following theorem by Brown *et al.* [13, Theorem 2.1].

THEOREM A. *Let $\mathbf{a} = a_0 a_1 a_2 \cdots$ be a k-automatic sequence over the alphabet Δ, and let $w \in \Delta^*$. Then the set of positions p such that w occurs beginning at position p is k-automatic.*

Concerning factors of length 1, the corresponding gap sequence is automatic too; this follows from [10].

The second part of our paper is concerned with the *discrepancy* of occurrences of $\mathtt{01}$-blocks in $\mathbf{t}$. More precisely, assume that N is a nonnegative integer. We count the number of times the factor $\mathtt{01}$ occurs in the first N terms of the Thue–Morse sequence, and compare it to $N/3$:

$$D_N := \#\{0 \le n < N : \mathbf{t}_n = \mathtt{0}, \mathbf{t}_{n+1} = 1\} - \frac{N}{3}. \tag{1-4}$$

From Theorem A we can immediately derive that the sequence $(D_N)_{N \ge 0}$ is 2-*regular* [2, 5] as the sequence of partial sums of a 2-automatic sequence: the sequence having

$$\begin{cases} 2/3 & \text{if } \mathbf{t}_n \mathbf{t}_{n+1} = \mathtt{01}, \\ -1/3 & \text{otherwise}, \end{cases}$$

as its nth term is automatic as the sum of four 2-automatic sequences, and D_N is the sum of the first N terms of this sequence [2, Theorem 3.1]. Our second theorem shows,

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

more specifically, that D_N can be obtained as the output sum of a base-2 transducer (see Heuberger *et al.* [18], in particular Remark 3.10 in that paper).

THEOREM 1.2. *The sequence $(D_N)_{N\ge 0}$ is the sequence of output sums of a base-2 transducer. In particular, $D_N \le C \log N$ for some absolute implied constant C. Moreover,*

$$\{D_N : N \ge 0\} = \tfrac{1}{3}\mathbb{Z}.$$

Note that the unboundedness of D_N follows from Corollary 4.10 in the paper by Berthé and Bernales [9] on balancedness in words.

1.1. Plan of the paper. In Section 2 we prove that the gap sequence for a factor w of $\mathbf{t}$ is not automatic. The central step of this proof is the case $w = \mathtt{01}$, which is handled in the first three subsections. Section 2.4 reduces the general case to this special case. In Section 3 we study the automatic sequence $\mathbf{A}$ on the three symbols $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, closely related to the gap sequences. In particular, we lift this sequence to the seven-letter alphabet $K = \{\mathtt{a}, \underline{\mathtt{b}}, \underline{\bar{\mathtt{b}}}, \underline{\mathtt{b}}, \underline{\mathtt{b}}, \underline{\mathtt{c}}, \underline{\mathtt{c}}\}$. From this new sequence we can in particular read off the discrepancy D_N easily, which leads to a proof of Theorem 1.2.

2. Proving the nonautomaticity of gap sequences

The main part of the proof of Theorem 1.1 concerns nonautomaticity of the gaps between occurrences of $\mathtt{01}$. As a second step in our proof, the general case is reduced to this one.

2.1. An auxiliary automatic sequence. We start by defining a substitution φ on three letters:

$$\varphi : \quad \mathtt{a} \mapsto \mathtt{abc}, \quad \mathtt{b} \mapsto \mathtt{ac}, \quad \mathtt{c} \mapsto \mathtt{b}. \tag{2-1}$$

The morphism φ can be extended to $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}^{\mathbb{N}}$ by concatenation, and we denote this extension by φ again. The unique fixed point (of length greater than 0) of φ is

$$\mathbf{A} = \mathtt{abcacbabcbacabcacbacabcbabcacbabcbacabcbabcacbac}\cdots.$$

This fixed point is a *morphic*, or *substitutive*, sequence [4, Ch. 7]. As a fixed point of φ (without having to apply a coding of the fixed point) it is even *pure morphic*. The sequence $\mathbf{A}$ is in fact 2-automatic, which follows from Berstel [7, Corollary 4]. It is a 'hidden automatic sequence', as treated very recently by Allouche *et al.* [1]. In fact, every automatic sequence can also be written as a coding of a fixed point of a nonuniform morphism [6] and in this sense is a 'hidden' automatic sequence. We restate a corresponding 2-uniform substitution found by Berstel in due course. The sequence $\mathbf{A}$, called the *ternary Thue–Morse sequence* (for example, in the On-Line Encyclopedia of Integer Sequences (OEIS) [26, A036577]), *Istrail squarefree sequence* [1, 19], or `vtm` [10], is well known. Citing Dekking [14], we note that it appears in fact 12 times on the OEIS [26], featuring all renamings of the letters

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

corresponding to permutations of the sets $\{0,1,2\}$ and $\{1,2,3\}$. These 12 entries are A005679, A007413, and A036577–A036586. The sequence $\mathbf{A}$ encodes the gaps between consecutive 1s in $\mathbf{t}$ [10]. Thue [27] showed that $\mathbf{A}$ is squarefree, while Rao *et al.* [24] later proved the stronger statement that $\mathbf{A}$ even avoids 2-*binomial squares* [25], thus settling in particular the question whether 2-abelian squares are avoidable over a three-letter alphabet. We use the squarefreeness property in our proof of Theorem 1.1.

LEMMA 2.1 (Thue). *The sequence* $\mathbf{A}$ *is squarefree. That is, no factor of the form* CC*, where* C *is a finite word over* $\{\mathtt{a},\mathtt{b},\mathtt{c}\}$ *of length at least* 1*, appears in* $\mathbf{A}$*.*

We have the following important relation between $\mathbf{A}$ and our problem.

LEMMA 2.2. *The Thue–Morse sequence* $\mathbf{t}$ *can be recovered from* $\mathbf{A}$ *via the substitution*

$$f : \mathtt{a} \mapsto \mathtt{011010}, \quad \mathtt{b} \mapsto \mathtt{0110}, \quad \mathtt{c} \mapsto \mathtt{01}, \tag{2-2}$$

by concatenation: we have

$$\mathbf{t} = f(\mathbf{A}_0)f(\mathbf{A}_1)\cdots. \tag{2-3}$$

We prove this in a moment. From this observation, noting also that each of the three words $f(\mathtt{a})$, $f(\mathtt{b})$, and $f(\mathtt{c})$ begins with $\mathtt{01}$, we see that we can extract from $\mathbf{A}$ the sequence of gaps between occurrences of the factor $\mathtt{01}$ in $\mathbf{t}$: each $\mathtt{a}$ yields two consecutive gaps of size 3, each $\mathtt{b}$ yields a gap of size 4, and each $\mathtt{c}$ a gap of size 2.

PROOF OF LEMMA 2.2. We prove that for $k \geq 1$,

$$\mathbf{t}_{[0,6\cdot 2^k]} = f(\mathbf{A}_0)f(\mathbf{A}_1)\cdots f(\mathbf{A}_{L_k}), \tag{2-4}$$

where $L_k = 3\cdot 2^{k-1} - 1$, by induction. The case $k = 1$ is just the trivial identity $\mathtt{011010011001} = f(\mathtt{a})f(\mathtt{b})f(\mathtt{c})$.

Note that $\tau(f(\mathtt{a})) = f(\mathtt{a})f(\mathtt{b})f(\mathtt{c})$, $\tau(f(\mathtt{b})) = f(\mathtt{a})f(\mathtt{c})$, and $\tau(f(\mathtt{c})) = f(\mathtt{b})$. Extending f by concatenation, for convenience of notation, to words over $\{\mathtt{a},\mathtt{b},\mathtt{c}\}$, we obtain $\tau(f(x)) = f(\varphi(x))$ for each $x \in \{\mathtt{a},\mathtt{b},\mathtt{c}\}$. An application of the morphism τ to both sides of (2-4) yields

$$\begin{aligned}\mathbf{t}_{[0,6\cdot 2^{k+1})} &= \tau(f(\mathbf{A}_0))\cdots\tau(f(\mathbf{A}_{L_k})) = f(\varphi(\mathbf{A}_0))\cdots f(\varphi(\mathbf{A}_{L_k}))\\ &= f(\varphi(\mathbf{A}_0\cdots\mathbf{A}_{L_k})) = f(\mathbf{A}_1\cdots\mathbf{A}_{L_{k+1}}) = f(\mathbf{A}_0)\cdots f(\mathbf{A}_{L_{k+1}}).\end{aligned}$$

Note that we see by induction that $\varphi^k(\mathtt{abc})$ contains each of the three letters 2^k times, hence the numbers L_k. This proves the lemma. □

Since each of the words $f(\mathtt{a})$, $f(\mathtt{b})$, and $f(\mathtt{c})$ starts with $\mathtt{01}$, the differences $a_j = k_{j+1} - k_j$ between successive occurrences of $\mathtt{01}$ in $\mathbf{t}$ are easily obtained from $\mathbf{A}$ by the substitution

$$r : \mathtt{a} \mapsto \mathtt{33}, \quad \mathtt{b} \mapsto \mathtt{4}, \quad \mathtt{c} \mapsto \mathtt{2}. \tag{2-5}$$

Here each $\mathtt{a}$ yields two blocks $\mathtt{01}$, and each $\mathtt{b}$ or $\mathtt{c}$ one block.

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

Let $\mathbf{B}$ be the sequence of gaps between consecutive occurrences of $\mathtt{01}$ in the Thue–Morse sequence, and $\check{\mathbf{B}}$ the corresponding sequence for $\mathtt{10}$.

LEMMA 2.3. *The sequence* $\mathbf{B}$ *is a morphic sequence, given by the substitution* ψ *on the four letters* $\mathtt{a}, \bar{\mathtt{a}}, \mathtt{b}, \mathtt{c}$*, together with the coding* p*, given by*

$$\begin{array}{lllll} \psi: & \mathtt{a} \mapsto \mathtt{a}\bar{\mathtt{a}}, & \bar{\mathtt{a}} \mapsto \mathtt{bc}, & \mathtt{b} \mapsto \mathtt{a}\bar{\mathtt{a}}\mathtt{c}, & \mathtt{c} \mapsto \mathtt{b}, \\ p: & \mathtt{a} \mapsto 3, & \bar{\mathtt{a}} \mapsto 3, & \mathtt{b} \mapsto 4, & \mathtt{c} \mapsto 2. \end{array} \tag{2-6}$$

Let $\bar{\mathbf{B}}$ *denote the fixed point of* ψ *starting with* $\mathtt{a}$.

The sequence $\check{\mathbf{B}}$ *is morphic. More precisely, it is the image of* $\bar{\mathbf{B}}$ *under the morphism*

$$\check{p}: \mathtt{a} \mapsto 24, \quad \bar{\mathtt{a}} \mapsto 33, \quad \mathtt{b} \mapsto 233, \quad \mathtt{c} \mapsto 4.$$

Note that $\bar{\mathbf{B}}$ is the pointwise limit of the finite words $\psi^k(\mathtt{a})$, and begins as follows:

$$\bar{\mathbf{B}} = \mathtt{a}\bar{\mathtt{a}}\mathtt{bca}\bar{\mathtt{a}}\mathtt{cba}\bar{\mathtt{a}}\mathtt{bca}\bar{\mathtt{a}}\mathtt{cba}\bar{\mathtt{a}}\mathtt{cba}\bar{\mathtt{a}}\mathtt{bca}\bar{\mathtt{a}}\mathtt{cba}\bar{\mathtt{a}}\mathtt{bca}\bar{\mathtt{a}}\mathtt{cba}\bar{\mathtt{a}}\mathtt{bcba}\bar{\mathtt{a}}\mathtt{c}\cdots.$$

PROOF. Let q be the morphism that replaces $\mathtt{a}$ by $\mathtt{a}\bar{\mathtt{a}}$ and leaves $\mathtt{b}$ and $\mathtt{c}$ unchanged. We show by induction on the length of a word C over $\mathtt{abc}$ that

$$q(\varphi(C)) = \psi(q(C)). \tag{2-7}$$

This is clear for words of length 1, since $q(\varphi(\mathtt{a})) = \mathtt{a}\bar{\mathtt{a}}\mathtt{bc} = \psi(q(\mathtt{a}))$, $q(\varphi(\mathtt{b})) = \mathtt{a}\bar{\mathtt{a}}\mathtt{c} = \psi(q(\mathtt{b}))$, and $q(\varphi(\mathtt{c})) = \mathtt{b} = \psi(q(\mathtt{c}))$. Appending a letter $x \in \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ to a word C for which the identity (2-7) already holds, we obtain

$$\begin{aligned} q(\varphi(Cx)) &= q(\varphi(C)\varphi(x)) = q(\varphi(C))q(\varphi(x)) \\ &= \psi(q(C))\psi(q(x)) = \psi(q(C)q(x)) = \psi(q(Cx)) \end{aligned}$$

and therefore (2-7) for C replaced by Cx. Next, we prove by induction, using (2-7), that

$$q(\varphi^k(\mathtt{a})) = \psi^k(q(\mathtt{a})).$$

Clearly, this holds for $k = 1$. For $k \geq 2$, we obtain

$$q(\varphi^k(\mathtt{a})) = q(\varphi(\varphi^{k-1}(\mathtt{a}))) = \psi(q(\varphi^{k-1}(\mathtt{a}))) = \psi(\psi^{k-1}(q(\mathtt{a}))) = \psi^k(q(\mathtt{a})).$$

Noting that $q(\mathtt{a}) = \psi(\mathtt{a})$ and $p \circ q = r$, the proof of the first part of Lemma 2.3 is complete.

We proceed to the second part, concerning $\check{\mathbf{B}}$. Note that by Corollary 7.7.5 in [4] we only have to prove that $\check{\mathbf{B}} = \check{p}(\bar{\mathbf{B}})$.

Let

$$\check{f}: \mathtt{a} \mapsto \mathtt{011010}, \quad \bar{\mathtt{a}} \mapsto \mathtt{011001}, \quad \mathtt{b} \mapsto \mathtt{01101001}, \quad \mathtt{c} \mapsto \mathtt{0110},$$

and extend this function to words (finite or infinite) over $\{\mathtt{a}, \bar{\mathtt{a}}, \mathtt{b}, \mathtt{c}\}$ by concatenation.

Applying τ, we see by direct computation that

$$\tau(\check{f}(\mathtt{a})) = \mathtt{011010011001} = \check{f}(\mathtt{a})\check{f}(\bar{\mathtt{a}}) = \check{f}(\psi(\mathtt{a})),$$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

and analogously, we get $\tau(\check{f}(x)) = \check{f}(\psi(x))$ for each letter $x \in \{\bar{\mathtt{a}}, \mathtt{b}, \mathtt{c}\}$. Applying this letter by letter, we obtain

$$\tau(\check{f}(w)) = \check{f}(\psi(w))$$

for every finite word over $\{\mathtt{a}, \bar{\mathtt{a}}, \mathtt{b}, \mathtt{c}\}$. By induction, we obtain

$$\tau^k(\check{f}(\mathtt{a})) = \check{f}(\psi^k(\mathtt{a})),$$

using the step

$$\tau^{k+1}(\check{f}(\mathtt{a})) = \tau(\tau^k(\check{f}(\mathtt{a}))) = \tau(\check{f}(\psi^k(\mathtt{a}))) = \check{f}(\psi^{k+1}(\mathtt{a})).$$

Noting that $\check{f}(\mathtt{a})$ begins with $\mathtt{0}$, we obtain $\mathbf{t} = \check{f}(\bar{\mathbf{B}})$. In other words, the sequence $\bar{\mathbf{B}}$ yields the decomposition $\mathbf{t} = x_0 x_1 \cdots$ of the Thue–Morse sequence into return words of $\mathtt{0110}$, where $x_j = \check{f}(\bar{\mathbf{B}}_j)$. From this decomposition we can easily read off the sequence of gaps between occurrences of $\mathtt{10}$, since this word appears in each of the four return words, and the first occurrence always takes place at the same position, which is 2. In this way, we obtain the gaps 2 and 3 from the return word $\check{f}(\mathtt{a})$ each time $\mathtt{a}$ appears in $\bar{\mathbf{B}}$. Analogously, $\bar{\mathtt{a}}$ yields the gaps 3 and 3, the letter $\mathtt{b}$ the gaps 2, 3, and 3, and finally $\mathtt{c}$ yields the gap 4. This proves the second part of Lemma 2.3. □

REMARK 2.4. A hint as to how to come up with the definition of ψ can be found by combining the substitutions φ and r, given in (2-1) and (2-5), respectively, and considering the first few words $w_k = r(\varphi^k(\mathtt{a}))$: we have $w_1 = \mathtt{3342}$, $w_2 = \mathtt{33423324}$, $w_3 = \mathtt{3342332433424332}$. We see that a first guess for a definition of ψ, choosing $\mathtt{3} \mapsto \mathtt{3342}$, leads to the incorrect result $\mathtt{33423342}\cdots$ after the next iteration; we are led to distinguishing between 'the first letter "3"' and 'the second letter "3"' in each occurrence of $\mathtt{33}$, which is exactly what our definition of ψ does. On the other hand, we directly obtain (2-6) by inspecting the decomposition of $\mathbf{t}$ into return words of $\mathtt{01}$. (Equivalently, we can study return words of $\mathtt{0110}$, as we did in the second part of the proof of Lemma 2.3.) We can write the image under τ of each return word as a concatenation of return words, which yields the desired morphism.

2.2. Factors of B appearing at positions in a residue class. The main step in our proof of Theorem 1.1 is given by the following proposition. For completeness, we let ψ^0 denote the identity, so that $\psi^0(w) = w$ for all words w over $\{\mathtt{a}, \bar{\mathtt{a}}, \mathtt{b}, \mathtt{c}\}$.

PROPOSITION 2.5. *Let $\mu \geq 0$ be an integer. The sequence of indices where $\psi^{4\mu}(\mathtt{a})$ appears as a factor in $\bar{\mathbf{B}}$ has nonempty intersection with every residue class $a + m\mathbb{Z}$, where $m \geq 1$ and a are integers.*

In the remainder of this section, we prove this proposition. We work with the fourth iteration $\sigma = \psi^4$ of the substitution ψ: we have

$$\begin{aligned} \sigma(\mathtt{a}) &= \mathtt{a\bar{a}bca\bar{a}cba\bar{a}bcbaac}, & \sigma(\bar{\mathtt{a}}) &= \mathtt{a\bar{a}bca\bar{a}cbaacaabcb}, \\ \sigma(\mathtt{b}) &= \mathtt{a\bar{a}bca\bar{a}cbaabcbaacaabcb}, & \sigma(\mathtt{c}) &= \mathtt{a\bar{a}bca\bar{a}cbaac}. \end{aligned} \tag{2-8}$$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

We have the following explicit formulas for the lengths of $\sigma^k(x)$, where $x \in \{\mathsf{a}, \bar{\mathsf{a}}, \mathsf{b}, \mathsf{c}\}$:

$$\begin{aligned} a_k &:= |\sigma^k(\mathsf{a})| = |\sigma^k(\bar{\mathsf{a}})| = 16^k, \\ b_k &:= |\sigma^k(\mathsf{b})| = \frac{4 \cdot 16^k - 1}{3}, \\ c_k &:= |\sigma^k(\mathsf{c})| = \frac{2 \cdot 16^k + 1}{3}. \end{aligned} \tag{2-9}$$

The proof of this identity is based on the formula

$$\begin{pmatrix} 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 \\ 5 & 5 & 6 & 5 \\ 3 & 3 & 2 & 3 \end{pmatrix}^k = \begin{pmatrix} 16^k/4 & 16^k/4 & 16^k/4 & 16^k/4 \\ 16^k/4 & 16^k/4 & 16^k/4 & 16^k/4 \\ (16^k-1)/3 & (16^k-1)/3 & (16^k+2)/3 & (16^k-1)/3 \\ (16^k+2)/6 & (16^k+2)/6 & (16^k-4)/6 & (16^k+2)/6 \end{pmatrix},$$

valid for $k \geq 1$, which takes care of the numbers of the letters a, $\bar{\mathsf{a}}$, b, and c in $\sigma^k(\mathsf{a})$, $\sigma^k(\bar{\mathsf{a}})$, $\sigma^k(\mathsf{b})$, and $\sigma^k(\mathsf{c})$. This formula can be proved easily by induction. Moreover, (2-9) also holds for $k = 0$.

By applying σ^k on the first line of (2-8), we see that each letter in $\{\mathsf{a}, \bar{\mathsf{a}}, \mathsf{b}, \mathsf{c}\}$ is replaced by a word having the respective lengths a_k, a_k, b_k, c_k. For each $\nu \geq 0$, it follows that the factor $\sigma^\nu(\mathsf{a})$, of length $a_\nu = 16^\nu$, can be found at the following positions in $\sigma^{\nu+1}(\mathsf{a})$:

$$\begin{aligned} A^{(\nu,0)} &:= 0, & A^{(\nu,1)} &:= 4 \cdot 16^\nu, \\ A^{(\nu,2)} &:= 8 \cdot 16^\nu, & A^{(\nu,3)} &:= 12 \cdot 16^\nu + \frac{4 \cdot 16^\nu - 1}{3}. \end{aligned} \tag{2-10}$$

We may repeat this for $\nu - 1, \nu - 2, \ldots, \mu$, where $\mu \leq \nu$ is a given natural number, from which we obtain the following statement. For all integers $0 \leq \mu \leq \nu$ and all $\boldsymbol{\varepsilon} = (\varepsilon_\mu, \varepsilon_{\mu+1}, \ldots, \varepsilon_\nu) \in \{0, 1, 2, 3\}^{\nu-\mu+1}$, the factor $\sigma^\mu(\mathsf{a})$ of length 16^μ can be found at the position

$$N_{\boldsymbol{\varepsilon}} := A^{(\mu,\varepsilon_\mu)} + A^{(\mu+1,\varepsilon_{\mu+1})} + \cdots + A^{(\nu,\varepsilon_\nu)} \tag{2-11}$$

in $\bar{\mathbf{B}}$. There are other positions where the factor $\sigma^\mu(\mathsf{a})$ appears, but for our proof it is sufficient to consider these special positions. We show that we can find one among these indices $N_{\boldsymbol{\varepsilon}}$ in a given residue class $a + m\mathbb{Z}$.

Let us sketch the remainder of the proof. The case where m is even causes mild difficulties. We therefore write $m = 2^k d$, where d is odd, and proceed in two steps. As a first step, we find integers μ, ν, and $\varepsilon_\mu, \varepsilon_{\mu+1}, \ldots, \varepsilon_{\lambda-1} \in \{0, 1, 2, 3\}$, such that $N_{\varepsilon_\mu, \varepsilon_{\mu+1}, \ldots, \varepsilon_{\lambda-1}}$ lies in any given residue class modulo 2^k. The second step involves refining the description by appending a sequence $(\varepsilon_\lambda, \ldots, \varepsilon_{\nu-1}) \in \{0, 1, 2\}^{\nu-\lambda}$. Since we exclude the digit $\varepsilon_i = 3$, and we take care that $16^\mu \geq 2^k$, we have

$$N_{\varepsilon_\mu, \ldots, \varepsilon_{\nu-1}} \equiv N_{\varepsilon_\mu, \ldots, \varepsilon_{\lambda-1}} \bmod 2^k.$$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

We choose the integers ε_j for $\lambda \le j < \mu$ in such a way that any given residue class modulo d (note that d is odd) is hit. Due to the excluded digit 3, this is a *missing digit* problem, and a short argument including exponential sums finishes this step. Combining these two steps, we see that every residue class modulo $2^k d$ is reached. We now go into the details.

First step: hitting a residue class modulo 2^k
We are interested in appearances of the initial segment $\sigma^\mu(\mathtt{a})$ in $\bar{\mathbf{B}}$ at positions lying in the residue class $a + 2^k\mathbb{Z}$. Let us assume in the following that

$$16^\mu \ge 2^k. \tag{2-12}$$

This lower bound on μ does not cause any problems.

We choose $\lambda > \mu$ in a moment, and we set $\varepsilon_\mu = \cdots = \varepsilon_{\lambda-1} = 3$. Let us consider the integers $\alpha_0 := 0$, and for $1 \le \ell \le \lambda - \mu$,

$$\alpha_\ell := N_{\varepsilon_\mu,\dots,\varepsilon_{\mu+\ell-1}}.$$

Assume that $0 \le \ell < \lambda - \mu$. By (2-11) and (2-12), we have

$$\begin{aligned}\alpha_{\ell+1} - \alpha_\ell &= 12 \cdot 16^{\mu+\ell} + \frac{4 \cdot 16^{\mu+\ell} - 1}{3} \equiv \frac{4 \cdot 16^{\mu+\ell} - 1}{3} \bmod 2^\ell \\ &\equiv \sum_{0 \le j \le 2\mu+2\ell} 4^j \bmod 2^\ell \equiv \sum_{0 \le j < 2\mu} 4^j \bmod 2^\ell.\end{aligned}$$

The latter sum is an odd integer, and independent of ℓ. It follows that $(\alpha_\ell)_{0 \le \ell \le \lambda-\mu}$ is an arithmetic progression modulo 2^k, where the common difference is odd; choosing $\lambda \ge \mu + 2^k$, we see that $(\alpha_\ell)_{0 \le \ell < \lambda-\mu}$ hits every residue class modulo 2^k. We summarize the first step in the following lemma.

LEMMA 2.6. *Let $k \ge 0$ and $\mu \ge k/4$, and choose $\varepsilon_{\mu+\ell} = 3$ for $\ell \ge 0$. The integers $N_{\varepsilon_\mu,\dots,\varepsilon_{\lambda-1}}$ hit every residue class modulo 2^k, as λ runs through the integers $\ge \mu$.*

Second step: a discrete Cantor set–missing digits
We follow the paper [17] by Erdős *et al.*, who studied integers with missing digits in residue classes. Let $\mathcal{W}_\lambda$ be the set of nonnegative multiples of 16^λ having only the digits 0, 4, and 8 in their base-16 expansion. Set

$$U(\alpha) = \frac{1}{3} \sum_{0 \le k \le 2} \mathrm{e}(4k\alpha) \quad \text{and} \quad G(\alpha, \lambda, \nu) = \frac{1}{3^{\nu-\lambda}} \sum_{\substack{0 \le j < 16^\nu \\ j \in \mathcal{W}}} \mathrm{e}(j\alpha),$$

where $\mathrm{e}(x) = \exp(2\pi i x)$. Note that the elements $j \in \mathcal{W}_\lambda$ have the form $j = \sum_{\lambda \le k < \eta} 4\,\varepsilon_k 16^k$, where $\eta \ge 0$ and $\varepsilon_k \in \{0, 1, 2\}$ for $\lambda \le k < \eta$. In particular,

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

$\mathcal{W}_\lambda \cap [0, 16^\eta)$ has $3^{\eta-\lambda}$ elements for $\eta \geq \lambda$. We obtain

$$\begin{aligned} G(\alpha, \lambda, \nu) &= \frac{1}{3^{\nu-\lambda}} \sum_{(\varepsilon_\lambda,\dots,\varepsilon_{\nu-1})\in\{1,2,3\}^{\nu-\lambda}} \mathrm{e}(4\,\varepsilon_\lambda 16^\lambda \alpha + \cdots + 4\,\varepsilon_{\nu-1} 16^{\nu-1}\alpha) \\ &= \prod_{\lambda\leq r<\nu} \tfrac{1}{3}(\mathrm{e}(0\cdot 16^r\alpha) + \mathrm{e}(4\cdot 16^r\alpha) + \mathrm{e}(8\cdot 16^r\alpha)) \\ &= \prod_{\lambda\leq r<\nu} U(16^r\alpha). \end{aligned} \tag{2-13}$$

The purpose of this section is to prove the following lemma.

LEMMA 2.7. *Let $\lambda \geq 0$ be an integer, and a, d integers such that $d \geq 1$ is odd. Then $\mathcal{W}_\lambda \cap (a + d\,\mathbb{Z})$ contains infinitely many elements.*

In order to prove this, we first show that it is sufficient to prove the following auxiliary result (compare [17, formula (4.3)]).

LEMMA 2.8. *Assume that $d \geq 1$ is an odd integer, and $\ell \in \{1, \dots, d-1\}$. Let $\lambda \geq 0$ be an integer. Then*

$$\lim_{\nu\to\infty} G\Big(\frac{\ell}{d}, \lambda, \nu\Big) = 0.$$

In fact, by the orthogonality relation

$$\frac{1}{d}\sum_{0\leq n<d} \mathrm{e}(nk/d) = \begin{cases} 1 & \text{if } d \mid k, \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\begin{aligned} &\frac{1}{3^{\nu-\lambda}}\#\{0 \leq j < 16^\nu : j \in \mathcal{W}, j \equiv a \bmod d\} - \frac{1}{d} \\ &\quad= \frac{1}{d}\frac{1}{3^{\nu-\lambda}} \sum_{0\leq\ell<d} \sum_{\substack{0\leq j<16^\nu \\ j\in\mathcal{W}}} \mathrm{e}(\ell(j-a)/d) - \frac{1}{d} \\ &\quad= \frac{1}{d}\sum_{1\leq\ell<d} \mathrm{e}(-\ell a/d)\frac{1}{3^{\nu-\lambda}} \sum_{\substack{0\leq j<16^\nu \\ j\in\mathcal{W}}} \mathrm{e}(j\ell/d) \leq \sum_{1\leq\ell<d} \Big|G\Big(\frac{\ell}{d}, \lambda, \nu\Big)\Big|. \end{aligned} \tag{2-14}$$

If $G(\ell/d, \lambda, \nu)$ converges to zero as $\nu \to \infty$, for all $\ell \in \{1, \dots, d-1\}$, the last sum in (2-14) is eventually smaller than $1/d$. Consequently, the number of $j \in \{0, \dots, 16^\nu - 1\}$ such that $j \in \mathcal{W}_\lambda$ and $j \equiv a \bmod d$ diverges to ∞ as ν approaches ∞.

PROOF OF LEMMA 2.8. By (2-14), we have to show that the product

$$\prod_{\lambda\leq r<\nu} U(16^r\ell/d) = \prod_{\lambda\leq r<\nu} (1 + \mathrm{e}(4\cdot 16^r\ell/d) + \mathrm{e}(8\cdot 16^r\ell/d)) \tag{2-15}$$

converges to zero as $\nu \to \infty$. To this end, we use the following lemma [15] by Delange.

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

LEMMA 2.9 (Delange). *Assume that $q \geq 2$ is an integer and $z_1, \dots, z_{q-1}$ are complex numbers such that $|z_j| \leq 1$ for $1 \leq j < q$. Then*

$$\left|\frac{1}{q}(1 + z_1 + \cdots + z_{q-1})\right| \leq 1 - \frac{1}{2q} \max_{1 \leq j < q}(1 - \operatorname{Re} z_j).$$

Since d is odd and $1 \leq \ell < d$, the integer $4k16^r\ell$ is not a multiple of d for $k \in \{1, 2\}$. It follows that $\operatorname{Re} e(4k16^r\ell/d) \leq 1 - \tilde{\varepsilon}$ for some $\tilde{\varepsilon} > 0$ only depending on d.

Therefore each factor in (2-13) is smaller than $1 - \varepsilon$, where $\varepsilon > 0$ does not depend on r. Consequently, by Lemma 2.9 the product (2-15) converges to zero. Lemma 2.8, and therefore Lemma 2.7, is proved. □

Now we combine the two steps, corresponding to the cases (i) 2^k and (ii) d odd.

Let $k \geq 0$ and $d \geq 1$ be integers, and d odd. We are interested in a residue class $a + 2^k d\mathbb{Z}$, where $a \in \mathbb{Z}$. Choose

$$a^{(1)} := a \bmod 2^k \in \{0, \dots, 2^k - 1\}.$$

Choose μ large enough such that $16^\mu \geq 2^k$. By Lemma 2.6 there exists $\lambda \geq \mu$ such that

$$\kappa^{(1)} \equiv a^{(1)} \bmod 2^k,$$

where $\kappa^{(1)} := N_{\varepsilon_\mu, \dots, \varepsilon_{\lambda-1}}$ and $\varepsilon_\ell = 3$ for $\mu \leq \ell < \lambda$. Next, choose

$$a^{(2)} := (a - \kappa^{(1)}) \bmod d.$$

By Lemma 2.7, the set $\mathcal{W}_\lambda \cap (a^{(2)} + d\mathbb{Z})$ is not empty. Let $\sum_{\lambda \leq \ell < \nu} 4\varepsilon_\ell 16^\ell$ be an element, where $\varepsilon_\ell \in \{0, 1, 2\}$ for $\lambda \leq \ell < \nu$. By (2-11) we have

$$\kappa := N_{\varepsilon_\mu, \dots, \varepsilon_{\lambda-1}, \varepsilon_\lambda, \dots, \varepsilon_{\nu-1}} = \kappa^{(1)} + \kappa^{(2)},$$

where

$$\kappa^{(2)} := N_{\varepsilon_\lambda, \dots, \varepsilon_{\nu-1}}.$$

The integer $\kappa^{(1)}$ lies in the residue class $a^{(1)} + 2^k\mathbb{Z}$ by construction, while $\kappa^{(2)}$ is divisible by 2^k, as no digit among $\varepsilon_\lambda, \dots, \varepsilon_{\nu-1}$ equals 3. It follows that $\kappa \in a^{(1)} + 2^k\mathbb{Z} = a + 2^k\mathbb{Z}$. Moreover, by (2-10),

$$\kappa^{(2)} = \sum_{\lambda \leq \ell < \nu} 4\varepsilon_\ell 16^\ell \in a^{(2)} + d\mathbb{Z},$$

hence $\kappa = \kappa^{(1)} + \kappa^{(2)} \equiv \kappa^{(1)} + (a - \kappa^{(1)}) \equiv a \bmod d$.

Summarizing, we have $\kappa \in (a + 2^k\mathbb{Z}) \cap (a + d\mathbb{Z})$. Since 2^k and d are coprime, which implies $2^k\mathbb{Z} \cap d\mathbb{Z} = 2^k d\mathbb{Z}$, we have $(a + 2^k\mathbb{Z}) \cap (a + d\mathbb{Z}) = a + 2^k d\mathbb{Z}$ (applying a shift by a) and therefore $\kappa \in a + 2^k d\mathbb{Z}$. This finishes the proof of Proposition 2.5. □

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

2.3. Nonautomaticity of B. In order to prove that **B** is not automatic, we use the characterization by the k-kernel: a sequence $(a_n)_{n\geq 0}$ is k-automatic if and only if the set

$$\{(a_{\ell+k^j n})_{n\geq 0} : j \geq 0, 0 \leq \ell < k^j\} \tag{2-16}$$

is finite.

We are now in a position to prove that *any* two arithmetic subsequences of **B** with the same modulus m and different shifts ℓ_1, ℓ_2 are different: the sequences $(\mathbf{B}(\ell_1 + nm))_{n\geq 0}$ and $(\mathbf{B}(\ell_2 + nm))_{n\geq 0}$ cannot be equal. This proves, in particular, that the k-kernel is infinite and thus nonautomaticity of the gap sequence for $\mathtt{01}$.

Let us assume, in order to obtain a contradiction, that the sequence **B** contains two identical arithmetic subsequences with common differences equal to m, indexed by $n \mapsto \ell_1 + nm$ and $n \mapsto \ell_2 + nm$ respectively, where $\ell_1 < \ell_2$. Let $r = \ell_2 - \ell_1$, and choose μ large enough such that $16^\mu \geq 2r$. By Proposition 2.5, the block $\sigma^\mu(\mathtt{a})$ appears in $\bar{\mathbf{B}}$ at positions that hit each residue class. In particular, for each $s \in \{0, \dots, r-1\}$, we choose the residue class $\ell_1 - s + m\mathbb{Z}$, and we can find an index n such that $\sigma^\mu(\mathtt{a})$ appears at position $\ell_1 - s + nm$ in $\bar{\mathbf{B}}$. Since $16^m \geq 2r > s$, this means that $\ell_1 + mn$ hits the sth letter in $\sigma^\mu(\mathtt{a})$, or in symbols,

$$\bar{\mathbf{B}}_{\ell_1+nm} = \sigma^\mu(\mathtt{a})|_s.$$

Since $s + r$ is still in the range $[0, 16^\mu)$, we also have

$$\bar{\mathbf{B}}_{\ell_2+nm} = \sigma^\mu(\mathtt{a})|_{s+r}$$

for the same index n. Applying the coding p defined in (2-6), and our equality assumption, we see that

$$\mathbf{B}_s = p(\sigma^\mu(\mathtt{a})|_s) = \mathbf{B}_{\ell_1+nm} = \mathbf{B}_{\ell_2+nm} = p(\sigma^\mu(\mathtt{a})|_{s+r}) = \mathbf{B}_{s+r}.$$

Carrying this out for all $s \in \{0, \dots, r-1\}$, we see that the first $2r$ terms of **B** form a square. Now there are two cases to consider.

The case $r = 1$. Assume that $\mathbf{B}_{\ell_1+nm} = \mathbf{B}_{\ell_1+1+nm}$ for all $n \geq 0$. By Proposition 2.5, the positions where the prefix $\mathtt{3342} = \mathbf{B}_0\mathbf{B}_1\mathbf{B}_2\mathbf{B}_3$ appears as a factor in **B** hit every residue class. In particular, there is an index n such that the block $\mathtt{3342}$ can be found at position $\ell_1 - 1 + nm$ in **B**. This implies $3 = \mathbf{B}_1 = \mathbf{B}_{\ell_1+nm} = \mathbf{B}_{\ell_1+1+nm} = \mathbf{B}_2 = 4$, a contradiction.

The case $r \geq 2$. In this case we resort to the fact, proved below, that **B** does not contain squares of length greater than 2. Therefore we get a contradiction also in this case. In order to complete the proof that **B** is not automatic, it remains to prove (the second part of) the following result.

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

LEMMA 2.10. *The infinite word* $\bar{\mathbf{B}}$ *is squarefree. The word* $\mathbf{B}$ *does not contain squares of length greater than* 2.

PROOF. We begin with the first statement. Note first that, by the morphism (2-5), letters '3' in $\mathbf{B}$ appear in pairs; moreover, the squarefreeness of $\mathbf{A}$ implies that there are no runs of three or more 3s. This implies that the morphism r defined in (2-5) can be 'reversed' in the sense that $\mathbf{A}$ can be restored from $\mathbf{B}$ by the (unambiguous) rule $\tilde{r} : 33 \mapsto \mathtt{a}, 4 \mapsto \mathtt{b}, 2 \mapsto \mathtt{c}$. Also, $\bar{\mathbf{B}}$ can be restored from $\mathbf{B}$ by the (unambiguous) rule $33 \mapsto \mathtt{a}\bar{\mathtt{a}}$, $4 \mapsto \mathtt{b}$, $2 \mapsto \mathtt{c}$, thus reversing the effect on $\bar{\mathbf{B}}$ of the morphism p defined in (2-6). In particular, since $\mathbf{A}$ is squarefree, each occurrence of the factor $\mathtt{a}\bar{\mathtt{a}}$ in $\bar{\mathbf{B}}$ is bordered by symbols from the set $\{\mathtt{b}, \mathtt{c}\}$ (where of course the first occurrence at 0 is not bordered on the left by another symbol).

Assume, in order to obtain a contradiction, that the square CC is a factor of $\bar{\mathbf{B}}$. We distinguish between two cases.

The case $|C| = 1$. Let C consist of a single symbol $x \in \{\mathtt{a}, \bar{\mathtt{a}}, \mathtt{b}, \mathtt{c}\}$. The squarefreeness of $\mathbf{A}$ forbids $x \in \{\mathtt{b}, \mathtt{c}\}$; moreover, we saw a moment ago that $\mathtt{a}$ and $\bar{\mathtt{a}}$ may only appear together, bordered by symbols $\in \{\mathtt{b}, \mathtt{c}\}$. This excludes the possibility $x \in \{\mathtt{a}, \bar{\mathtt{a}}\}$, therefore this case leads to a contradiction.

The case $|C| \geq 2$. There are two subcases to consider. (i) Assume that C begins with $\bar{\mathtt{a}}$. In this case, C has to end with $\mathtt{a}$: the concatenation CC has to be a factor of $\bar{\mathbf{B}}$, and therefore the symbol $\bar{\mathtt{a}}$ at the start of the second 'C' has to be preceded by a symbol $\mathtt{a}$. Analogously, each occurrence of the word CC is immediately preceded by $\mathtt{a}$, and followed by $\bar{\mathtt{a}}$. That is, $\mathtt{a}CC\bar{\mathtt{a}}$ appears as a factor of $\bar{\mathbf{B}}$. Writing $C = \bar{\mathtt{a}}y\mathtt{a}$ for a finite (possibly empty) word y over $\{\mathtt{a}, \bar{\mathtt{a}}, \mathtt{b}, \mathtt{c}\}$, we see that $\mathtt{a}\bar{\mathtt{a}}y\mathtt{a}\bar{\mathtt{a}}y\mathtt{a}\bar{\mathtt{a}}$ is a factor of $\bar{\mathbf{B}}$. Applying the coding p, it follows that $T = \mathtt{aa}y\mathtt{aa}y\mathtt{aa}$ appears in $\mathbf{B}$, and it is a concatenation of the words 33, 4, and 2. Consequently, it makes sense to apply the 'inverse morphism' $\tilde{r} : 33 \mapsto \mathtt{a}$, $4 \mapsto \mathtt{b}$, $2 \mapsto \mathtt{c}$. Therefore $\tilde{r}(T) = \mathtt{a}z\mathtt{a}z\mathtt{a}$, for some finite word z over $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, appears in $\mathbf{A}$. This contradicts Lemma 2.1. (ii) Assume that C starts with a letter $\in \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. In this case, C ends with a letter $\in \{\bar{\mathtt{a}}, \mathtt{b}, \mathtt{c}\}$; otherwise, the concatenation CC, and therefore $\bar{\mathbf{B}}$, would contain $\mathtt{aa}$, which we have already ruled out. We apply p, and in this case $p(C)$ is a concatenation of the words 33, 4, and 2. Therefore we can form $\tilde{r}(p(C))$, revealing that the square $\tilde{r}(p(C))\tilde{r}(p(C))$ is a factor of $\mathbf{A}$. This is a contradiction.

We have to prove the second statement. Assume that CC is a factor of $\mathbf{B}$, where $|C| \geq 2$. This proof is analogous to the corresponding case for $\bar{\mathbf{B}}$, and we skip some of the details that we have already seen there. (i) Assume that C begins with exactly one $\mathtt{a}$. In this case, C has to end with exactly one $\mathtt{a}$, and therefore $C = \mathtt{a}y\mathtt{a}$ for a finite word y over $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. It follows that $\mathtt{aa}y\mathtt{aa}y\mathtt{aa}$ is a factor of $\mathbf{B}$. Applying $\tilde{r}$, we obtain a contradiction to Lemma 2.1.

(ii) Assume that C starts with $\mathtt{aa}$, $\mathtt{b}$, or $\mathtt{c}$. In this case, C ends with $\mathtt{aa}$, $\mathtt{b}$, or $\mathtt{c}$, otherwise CC, and therefore $\bar{\mathbf{B}}$, would contain a block of as of length not equal to 2.

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

We apply p on the word CC, followed by $\tilde{r}$, which yields the square $\tilde{r}(p(C))\tilde{r}(p(C))$. Again, this contradicts Lemma 2.1. □

Summarizing, arithmetic subsequences of $\mathbf{B}$ with common difference m are distinct as soon as their offsets differ. In particular, for each integer $k \geq 2$, the k-kernel of $\mathbf{B}$ is infinite. Therefore $\mathbf{B}$ is not automatic, which proves the case $w = \mathtt{01}$ of Theorem 1.1.

2.4. Occurrences of general factors in t. We begin with the case $w = \mathtt{10}$. We work with the Thue–Morse morphism $\tau : \mathtt{0} \mapsto \mathtt{01}$, $1 \mapsto \mathtt{10}$, defined in (1-1). First of all, we recall the well-known fact that $a_{k+1} = \tau^{k+1}(\mathtt{0})$ can be constructed from $a_k = \tau^k(\mathtt{0})$ by concatenating a_k and its Boolean complement $\overline{a_k}$ (which replaces each $\mathtt{0}$ by 1 and each 1 by $\mathtt{0}$). The proof of this little fact is by an easy induction. For $k = 0$ we have $a_1 = \mathtt{01} = \mathtt{0}\overline{\mathtt{0}}$. The case $k \geq 1$ makes use of the identity $\tau(\overline{w}) = \overline{\tau(w)}$, valid for each word w over $\{\mathtt{0}, 1\}$, which follows from the special structure of the morphism τ. Applying this identity and the induction hypothesis, we obtain

$$\begin{aligned} a_{k+1} &= \tau(\tau^k(\mathtt{0})) = \tau(a_{k-1}\overline{a_{k-1}}) \\ &= \tau(a_{k-1})\tau(\overline{a_{k-1}}) = \tau(a_{k-1})\overline{\tau(a_{k-1})} = a_k\overline{a_k}. \end{aligned}$$

Using this, we show that for even $k \geq 0$, the word $\tau^k(\mathtt{0})$ is a palindrome. The case $k = 0$ is trivial. If $a_k = \tau^k(\mathtt{0})$ is a palindrome, then $a_{k+2} = \tau(\tau(a_k)) = \tau(a_k\overline{a_k}) = a_k\overline{a_k}\overline{a_k}a_k$ is clearly a palindrome too, and the statement follows by induction. In particular, we see from the above that

$$\tau^k(\mathtt{0}) = \tau^{k-1}(\mathtt{0})\tau^{k-1}(1), \quad \tau^k(1) = \tau^{k-1}(1)\tau^{k-1}(\mathtt{0}), \quad \text{for all } k \geq 0. \tag{2-17}$$

Note that, by applying τ^k on $\mathbf{t}$, every $\mathtt{0}$ gets replaced by $\tau^k(\mathtt{0})$ and every 1 by $\tau^k(1)$, and the result is again $\mathbf{t}$ since it is a fixed point of τ. It follows that

$$\begin{aligned} &\text{for all } k \geq 0, \quad \text{we have } \mathbf{t} = A_{\mathbf{t}_0}A_{\mathbf{t}_1}A_{\mathbf{t}_2}\cdots, \\ &\qquad\qquad\quad \text{where } A_x = \tau^k(x) \text{ for } x \in \{\mathtt{0}, 1\}. \end{aligned} \tag{2-18}$$

Let $(r_j)_{j\geq 0}$ be the increasing sequence of indices where $\mathtt{10}$ occurs in $\mathbf{t}$. For k even, let $J = J(k)$ be the number of occurrences of $\mathtt{10}$ with indices less than or equal to $2^k - 2$. Note that $r_{J-1} = 2^k - 2$. We read the (palindromic) sequence a_k, of length 2^k, backwards; it follows that $(2^k - 1 - r_{J-1-j})_{0\leq j<J}$ is the increasing sequence of indices pointing to the letter 1 in an occurrence of $\mathtt{01}$ in a_k. Therefore,

$$(2^k - 2 - r_{J-1-j})_{0\leq j<J}$$

is the increasing sequence of indices where $\mathtt{01}$ occurs in a_k. Consequently, by the definition of $\mathbf{B}$ as the differences of these indices, we obtain $\mathbf{B}_j = -r_{J-1-(j+1)} + r_{J-1-j}$ for $0 \leq j < J - 1$, and thus

$$r_{j+1} - r_j = \mathbf{B}_{J-2-j}, \quad \text{for } 0 \leq j \leq J - 2. \tag{2-19}$$

We have to prove that the sequence

$$\check{\mathbf{B}} = (r_{j+1} - r_j)_{j\geq 0}$$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

is not automatic. More generally, we prove that any two arithmetic subsequences

$$L^{(1)} = (\check{\mathbf{B}}(\ell_1 + nd))_{n\geq 0}, \quad L^{(2)} = (\check{\mathbf{B}}(\ell_2 + nd))_{n\geq 0},$$

where $d \geq 1$ and $\ell_1 \neq \ell_2$, are different. In order to obtain a contradiction, let us assume that $L^{(1)} = L^{(2)}$, and let $k \geq 0$ be even. By (2-19), we get arithmetic subsequences M_1, M_2 of $\mathbf{B}$ with common difference d, different offsets $m_1(k), m_2(k) \in \{0, \dots, d-1\}$ and length equal to $J(k) - 1$, such that

$$\mathbf{B}_{m_1(k)+nd} = M_j^{(1)} = M_j^{(2)} = \mathbf{B}_{m_2(k)+nd}, \quad \text{for } 0 \leq n \leq J(k) - 2.$$

Note the important fact that the offsets $m_j(k)$ are bounded by d. Since there are only $d(d-1)/2$ pairs $(a, b) \in \{0, \dots, d-1\}^2$ with $a \neq b$, it follows that there are two different offsets $0 \leq \overline{m_1}, \overline{m_2} < d$ with the following property: there are arbitrarily long arithmetic subsequences of $\mathbf{B}$ with indices of the form $\overline{m_1} + nd$ and $\overline{m_2} + nd$ respectively, taking the same values. This is just the statement that the *infinite* sequences $(\mathbf{B}_{\overline{m_1}+nd})_{n\geq 0}$ and $(\mathbf{B}_{\overline{m_2}+nd})_{n\geq 0}$ are equal. In the course of proving that $\mathbf{B}$ is not automatic (which is the case $w = \mathtt{01}$ of Theorem 1.1) we proved that this is impossible, and we obtain a contradiction. The sequence $\check{\mathbf{B}}$ is therefore not automatic either, which finishes the case $w = \mathtt{10}$.

We proceed to the case $w = \mathtt{00}$. Let $(a_i)_{i\geq 0}$ be the increasing sequence of indices j such that $\mathbf{t}_j\mathbf{t}_{j+1} = \mathtt{00}$. Assume that $i \geq 0$, and set $j := a_i$. We have $j \equiv 1 \bmod 2$, since $\mathbf{t}_{2j'} = \overline{\mathbf{t}_{2j'+1}}$ for all $j' \geq 0$ (where the overline denotes the Boolean complement, $\mathtt{0} \mapsto 1$, $1 \mapsto \mathtt{0}$). Equality $\mathbf{t}_j = \mathbf{t}_{j+1}$ (as needed) can therefore only occur at odd indices j, and we choose $j' \geq 0$ such that $j = 2j' + 1$. Necessarily, $\mathbf{t}_{j'} = 1$ and $\mathbf{t}_{j'+1} = \mathtt{0}$, since the identities $\mathbf{t}_{2j'+1} = \overline{\mathbf{t}_{j'}}$ and $\mathbf{t}_{2j'+2} = \mathbf{t}_{j'+1}$ would produce an output $\mathbf{t}_{2j'+1}\mathbf{t}_{2j'+2} \neq \mathtt{00}$ in the other case. On the other hand, $\mathbf{t}_{j'}\mathbf{t}_{j'+1} = \mathtt{10}$ indeed implies $\mathbf{t}_{2j'+1}\mathbf{t}_{2j'+2} = \mathtt{00}$. Each occurrence of $\mathtt{00}$ in $\mathbf{t}$, at position j, therefore corresponds in a bijective manner to an occurrence of $\mathtt{10}$ at position $(j-1)/2$ (which is an integer). It follows that the corresponding gap sequence equals $2\check{\mathbf{B}}$, which is not automatic by the already proved case $w = \mathtt{10}$.

In a completely analogous manner, we can reduce the case $w = 11$ to the case $\mathtt{01}$, and the gap sequence equals $2\mathbf{B}$, which is not automatic either.

We now reduce the case of general factors w of $\mathbf{t}$ of length at least 3 to these four cases.

LEMMA 2.11. *For $x, y \in \{\mathtt{0}, 1\}$, let $(a_k^{xy})_{k\geq 0}$ be the increasing sequence of indices j such that $\mathbf{t}_j\mathbf{t}_{j+1} = xy$. We have*

$$\begin{aligned} &a_0^{\mathtt{01}} < a_0^{\mathtt{10}} < a_1^{\mathtt{01}} < a_1^{\mathtt{10}} < a_2^{\mathtt{01}} < a_2^{\mathtt{10}} < \cdots \quad \text{and} \\ &a_0^{11} < a_0^{\mathtt{00}} < a_1^{11} < a_1^{\mathtt{00}} < a_2^{11} < a_2^{\mathtt{00}} < \cdots . \end{aligned} \tag{2-20}$$

PROOF. First of all, $\mathbf{t}$ begins with $\mathtt{011}$, which explains the first item in each of the two displayed chains of inequalities. The first chain is almost trivial since after each block of consecutive $\mathtt{0}$s, a letter 1 follows, and vice versa.

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

Let us prove the second series of inequalities by induction. Assume that $a_0^{(11)} < a_0^{(00)} < \cdots < a_{i-1}^{(00)} < a_i^{(11)} = j$. Then $\mathbf{t}_j\mathbf{t}_{j+1} = 11$, and it follows that $\mathbf{t}_{j+2} = 0$, since 111 is not a factor of $\mathbf{t}$. Two cases can occur. (i) If $\mathbf{t}_{j+3} = 0$, then clearly $a_i^{(11)} < a_i^{(00)} = j + 2$ by our hypothesis. (ii) Otherwise, we have $\mathbf{t}_j\mathbf{t}_{j+1}\mathbf{t}_{j+2}\mathbf{t}_{j+3} = 1101$. Necessarily, j is odd: if $j = 2j'$, it would follow that $\mathbf{t}_j\mathbf{t}_{j+1} \in \{01, 10\}$, but we need 11. Moreover, $j \equiv 3 \bmod 4$ is also not possible. Let $j + 1 = 4j'$. Then $\mathbf{t}_{j+1}\mathbf{t}_{j+2}\mathbf{t}_{j+3} \in \{011, 100\}$, but we need 101. It follows that $j \equiv 1 \bmod 4$, and therefore $\mathbf{t}_{j+4}\mathbf{t}_{j+5} = 00$, which implies $a_i^{(00)} = j + 4$. By a completely analogous argument (reversing the roles of 1 and 0), we may finish the proof of Lemma 2.11 by induction. □

Let w be a factor of $\mathbf{t}$, of length at least 3. Choose $k \geq 0$ minimal such that w is a factor of some $a_k^{xy} = \tau^k(x)\tau^k(y)$, where $x, y \in \{0, 1\}$. By minimality, w is not a factor of $\tau^k(0)$ or $\tau^k(1)$, using (2-17). Consequently, w appears at most once in each a_k^{xy}. Next, we need the fact that $\mathbf{t}$ is *overlap-free* [8, 13, 27], meaning that it does not contain a factor of the form $axaxa$, where $a \in \{0, 1\}$ and $x \in \{0, 1\}^*$. We derive from this property that w cannot occur simultaneously in both members of any of the pairs

$$(a_k^{00}, a_k^{01}), \quad (a_k^{00}, a_k^{10}), \quad (a_k^{11}, a_k^{01}), \quad (a_k^{11}, a_k^{10}). \tag{2-21}$$

For example, assume that w is a factor of both a_k^{00} and a_k^{01}. By minimality, as we had before,

$$\tau^k(0)\tau^k(0) = AwB, \quad \tau^k(0)\tau^k(1) = A'wB',$$

where A and A' are initial segments of $\tau^k(0)$, and B (respectively, B') are final segments of $\tau^k(0)$ (respectively, $\tau^k(1)$), and all of these segments are proper subwords of the respective words. We have $A \neq A'$, since otherwise $\tau^k(0) = \tilde{w}B = \tilde{w}B' = \overline{\tau^k(1)}$ for some $\tilde{w}$ that is not the empty word. This contradicts the fact that $\tau^k(0) = \overline{\tau^k(1)}$. Let us, without loss of generality, assume that $|A| < |A'|$. The first 2^k letters of Aw and $A'w$ are equal, or in symbols,

$$(Aw)|_{[0,2^k)} = (A'w)|_{[0,2^k)}. \tag{2-22}$$

We can therefore choose $a \in \{0, 1\}$ and $w_1, w_2 \in \{0, 1\}^*$ in such a way that $aw_1w_2 = w$ and $Aaw_1 = A'$. Then trivially $Aw = Aaw_1w_2 = A'w_2$, and since $|A| < 2^k$, $|A'| < 2^k$, it follows from (2-22) that $w_2 = aw_3$ for some $w_3 \in \{0, 1\}^*$. Finally, the factor $A'w$ of $\mathbf{t}$ can be written as $A'w = Aaw_1w = Aaw_1aw_1w_2 = Aaw_1aw_1aw_3$, which contradicts the overlap-freeness of $\mathbf{t}$. The other three cases, corresponding to the second, third and fourth pairs in (2-21), are analogous. We have therefore shown that the set of $A \in \{a_k^{00}, a_k^{01}, a_k^{10}, a_k^{11}\}$ such that w is a factor of A is a subset of either $\{a_k^{01}, a_k^{10}\}$ or $\{a_k^{00}, a_k^{11}\}$.

First case. Let w be a factor of a_k^{01}, or of a_k^{10}. Assume first that w is a factor of a_k^{01}, but not of a_k^{10}. In this case, we show that the gap sequence for w is given by the gap sequence for a_k^{01}: (i) each occurrence of a_k^{01} yields exactly one occurrence

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

of w (involving a constant shift); (ii) by (2-18), every occurrence of w takes place within a block of the form a_k^{xy}; (iii) only the block $a_k^{\mathtt{01}}$ is eligible. We prove that $a_k^{\mathtt{01}}$ appears exactly at positions $2^k j$ in $\mathbf{t}$, where $\mathbf{t}_j\mathbf{t}_{j+1} = \mathtt{01}$. The easy direction follows from (2-18): each occurrence of $\mathtt{01}$ yields an occurrence of $a_k^{\mathtt{01}}$, where the index has to be multiplied by 2^k. On the other hand, it is sufficient to show that $a_k^{\mathtt{01}}$ can only appear in positions $2^k j$. Given this, there is no admissible choice for $(\mathbf{t}_j, \mathbf{t}_{j+1})$ different from $(\mathtt{0}, 1)$, by (2-18). Suppose that we already know this for some $k \geq 0$ (the case $k = 0$ being trivial). Assume that

$$a_{k+1}^{\mathtt{01}} = \tau^k(\mathtt{0})\tau^k(1)\tau^k(1)\tau^k(\mathtt{0}) \text{ appears on some position } \ell. \tag{2-23}$$

Since $\tau^k(\mathtt{0})\tau^k(1) = a_k^{\mathtt{01}}$, we know by hypothesis that $\ell \equiv 0 \bmod 2^k$. Assume that the case $\ell \equiv 2^k \bmod 2^{k+1}$ occurs. We set $\ell = (2j+1)2^k$ for some $j \geq 0$. Our assumption (2-23) implies $\tau^k(1) = \tau^k(\mathbf{t}_{2j+2}) = \tau^k(\mathbf{t}_{j+1})$ and therefore $\mathbf{t}_{j+1} = 1$, which implies that $\tau^{k+1}(\mathbf{t}_{j+1}) = \tau^k(1)\tau^k(\mathtt{0})$ appears in position $\ell + 2^k = (2j+2)2^k$ in $\mathbf{t}$. This is incompatible with (2-23). In particular, the gap sequence for w, which is identical to the gap sequence for $a_k^{\mathtt{01}}$, is given by $2^k\mathbf{B}$, and therefore not automatic. Switching the roles of $\mathtt{0}$ and 1 in this proof, we also obtain nonautomaticity for the case where w is a factor of $a_k^{\mathtt{10}}$, but not of $a_k^{\mathtt{01}}$, with the sequence $2^k\check{\mathbf{B}}$ as the corresponding gap sequence.

Let w be a factor of both $a_k^{\mathtt{01}}$ and $a_k^{\mathtt{10}}$. In this case, each occurrence of w in $\mathbf{t}$ takes place within a subblock of $\mathbf{t}$ of one of these two forms. By Lemma 2.11, combined with the above argument that occurrences of $a_k^{\mathtt{01}}$ (respectively, $a_k^{\mathtt{10}}$) in $\mathbf{t}$ take place at indices obtained from occurrences of $\mathtt{01}$ (respectively, $\mathtt{10}$), multiplied by 2^k, these blocks occur alternatingly. Assuming, in order to obtain a contradiction, that the gap sequence $(g_j)_{j\geq 0}$ for w is automatic, we obtain a new automatic sequence $(g_{2j} + g_{2j+1})_{j\geq 0}$ as the sum of two automatic sequences (note that the characterization involving the 2-kernel (2-16) immediately implies that $(g_{2j+\varepsilon})_{j\geq 0}$, for $\varepsilon \in \{0, 1\}$, is automatic). By the alternating property, this is the gap sequence for $a_k^{\mathtt{01}}$, which is not automatic, as we have just seen. A contradiction!

Second case. Let w be a factor of $a_k^{\mathtt{00}}$ or of a_k^{11}. This case is largely analogous. Assume that w is not a factor of a_k^{11}. As in the case $a_k^{\mathtt{01}}$, the gap sequence for w in this case is identical to the gap sequence for $a_k^{\mathtt{00}}$, and we only have to show that this sequence is not automatic. We know already that the gap sequence for $\mathtt{00}$ is not automatic. Therefore it suffices to prove that $\tau^k(\mathtt{0})\tau^k(\mathtt{0})$ can only appear at positions in $\mathbf{t}$ divisible by 2^k. Suppose that we already know this for some $k \geq 0$ (the case $k = 0$ again being trivial). Assume that

$$a_{k+1}^{\mathtt{00}} = \tau^k(\mathtt{0})\tau^k(1)\tau^k(\mathtt{0})\tau^k(1) \text{ appears on some position } \ell. \tag{2-24}$$

Since $\tau^k(\mathtt{0})\tau^k(1) = a_k^{\mathtt{01}}$, we know by hypothesis that $\ell \equiv 0 \bmod 2^k$. Assume that the case $\ell \equiv 2^k \bmod 2^{k+1}$ occurs. We set $\ell = (2j+1)2^k$ for some $j \geq 0$. Our assumption (2-24) implies $\tau^k(1) = \tau^k(\mathbf{t}_{2j+4}) = \tau^k(\mathbf{t}_{j+2})$ and therefore $\mathbf{t}_{j+2} = 1$, which implies that $\tau^{k+1}(\mathbf{t}_{j+2}) = \tau^k(1)\tau^k(\mathtt{0})$ appears in position $\ell + 3 \cdot 2^k = (2j+4)2^k$ in $\mathbf{t}$. In position ℓ, we

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

therefore see the factor

$$\tau^k(\mathtt{0})\tau^k(1)\tau^k(\mathtt{0})\tau^k(1)\tau^k(\mathtt{0}),$$

which contradicts the overlap-freeness of $\mathbf{t}$.

Again, the case where w is a factor of a_k^{11}, but not of $a_k^{\mathtt{00}}$, is analogous; the case that it is a factor of both words can be handled as in the case $\{a_k^{\mathtt{01}}, a_k^{\mathtt{10}}\}$, this time with the help of the second chain of inequalities in (2-20).

Summarizing, we have shown nonautomaticity for all gap sequences for factors w of $\mathbf{t}$ of length at least 2.

In order to finish the proof of Theorem 1.1, we still have to prove that the gap sequence is morphic for the 'mixed cases'. That is, assume that w is a factor of two words of the form a_k^{xy}, where $x, y \in \{\mathtt{0}, 1\}$, and where k is chosen minimal such that w is a factor of at least one of $a_k^{\mathtt{00}}, a_k^{\mathtt{01}}, a_k^{\mathtt{10}}, a_k^{11}$. Let us begin with the case $\{a_k^{\mathtt{01}}, a_k^{\mathtt{10}}\}$. The positions where w appears in $\mathbf{t}$ are given by $2^k j + \sigma_0$, where $\mathtt{b}_j\mathtt{b}_{j+1} = \mathtt{01}$, and $2^k j + \sigma_1$, where $\mathtt{b}_j\mathtt{b}_{j+1} = \mathtt{10}$. Here σ_0, σ_1 are the positions where the word w appears in $a_k^{\mathtt{01}}$ and $a_k^{\mathtt{10}}$, respectively. As before, this follows since $\mathbf{t}_j\mathbf{t}_{j+1} = \mathtt{01}$ is equivalent to $(\mathbf{t}_\ell, \dots, \mathbf{t}_{\ell+2^{k+1}-1}) = a_k^{\mathtt{01}}$, and from the corresponding statement for $\mathtt{10}$. We see that it is sufficient to write $\mathbf{t}$ as a concatenation of the words

$$w_{\mathtt{a}} := \mathtt{011}, \quad w_{\bar{\mathtt{a}}} := \mathtt{010}, \quad w_{\mathtt{b}} := \mathtt{0110}, \quad \text{and} \quad w_{\mathtt{c}} := \mathtt{01}, \tag{2-25}$$

since each word w_x takes care of one $\mathtt{01}$-block, followed by one $\mathtt{10}$-block, and the gap sequence for w is obtained by replacing each w_x by a succession of two gaps. Applying the morphism τ, we obtain $\tau(w_{\mathtt{a}}) = w_{\mathtt{a}}w_{\bar{\mathtt{a}}}$, $\tau(w_{\bar{\mathtt{a}}}) = w_{\mathtt{b}}w_{\mathtt{c}}$, $\tau(w_{\mathtt{b}}) = w_{\mathtt{a}}w_{\bar{\mathtt{a}}}w_{\mathtt{c}}$, $\tau(w_{\mathtt{c}}) = w_{\mathtt{b}}$. This mimics the morphism ψ; proceeding as in the proof of Lemma 2.2 (alternatively, as in the proof of Lemma 2.3), we obtain

$$\mathbf{t} = w_{\bar{\mathbf{B}}_0} w_{\bar{\mathbf{B}}_1} w_{\bar{\mathbf{B}}_2} \cdots . \tag{2-26}$$

Since $\bar{\mathbf{B}}$ is morphic, the succession of gaps with which w occurs in $\mathbf{t}$ is morphic by [4, Corollary 7.7.5] (that is, 'morphic images of morphic sequences are morphic').

The case $\{a_k^{\mathtt{00}}, a_k^{11}\}$ is similar. Defining

$$\tilde{w}_{\mathtt{a}} := \mathtt{011010}, \quad \tilde{w}_{\bar{\mathtt{a}}} := \mathtt{011001}, \quad \tilde{w}_{\mathtt{b}} := \mathtt{01101001}, \quad \text{and} \quad \tilde{w}_{\mathtt{c}} := \mathtt{0110},$$

it is straightforward to verify that $\tau(\tilde{w}_{\mathtt{a}}) = \tilde{w}_{\mathtt{a}}\tilde{w}_{\bar{\mathtt{a}}}$, $\tau(\tilde{w}_{\bar{\mathtt{a}}}) = \tilde{w}_{\mathtt{b}}\tilde{w}_{\mathtt{c}}$, $\tau(\tilde{w}_{\mathtt{b}}) = \tilde{w}_{\mathtt{a}}\tilde{w}_{\bar{\mathtt{a}}}\tilde{w}_{\mathtt{c}}$, and $\tau(\tilde{w}_{\mathtt{c}}) = \tilde{w}_{\mathtt{b}}$. Again, we can spot the morphism ψ, and we obtain

$$\mathbf{t} = \tilde{w}_{\bar{\mathbf{B}}_0} \tilde{w}_{\bar{\mathbf{B}}_1} \tilde{w}_{\bar{\mathbf{B}}_2} \cdots$$

in exactly the same way as before. Each of the words w_x in this representation yields a block 11 in $\mathbf{t}$, followed by a block $\mathtt{00}$. Therefore, also in this case, the gap sequence for w is a morphic image of a morphic sequence. This finishes the proof of Theorem 1.1.

REMARK 2.12. Let us have a closer look at the gaps in the 'mixed case' $\{a_k^{\mathtt{01}}, a_k^{\mathtt{10}}\}$. Let σ_0 be the index at which w appears in $a_k^{\mathtt{01}}$, and σ_1 the index at which w appears in $a_k^{\mathtt{10}}$. By (2-26) and the choice (2-25), each letter $x \in \{\mathtt{a}, \bar{\mathtt{a}}, \mathtt{b}, \mathtt{c}\}$ in $\bar{\mathbf{B}}$ corresponds to

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

two gaps, as follows.

Letter in $\bar{\mathbf{B}}$	Gap 1	Gap 2
a	$\sigma_1 - \sigma_0 + 2^{k+1}$	$\sigma_0 - \sigma_1 + 2^k$
ā	$\sigma_1 - \sigma_0 + 2^k$	$\sigma_0 - \sigma_1 + 2^{k+1}$
b	$\sigma_1 - \sigma_0 + 2^{k+1}$	$\sigma_0 - \sigma_1 + 2^{k+1}$
c	$\sigma_1 - \sigma_0 + 2^k$	$\sigma_0 - \sigma_1 + 2^k$

It follows that there are at most four gaps that can occur in this case. For example, consider the gap sequence for the factor $w = \mathtt{010}$. In this case $k = 2$, and we have $a_2^{\mathtt{01}} = \mathtt{011\underline{010}01}$ and $a_2^{\mathtt{10}} = \mathtt{100\underline{10}110}$, where the occurrences of w are underlined. We have $\sigma_0 = 3$ and $\sigma_1 = 2$. This yields the gaps 3, 5, 7, and 9, occurring only in the combinations $(7,5)$, $(3,9)$, $(7,9)$, and $(3,5)$. Noting also the first occurrence $\mathbf{t}_3\mathbf{t}_4\mathbf{t}_5 = \mathtt{010}$, the first few occurrences of $\mathtt{010}$ in $\mathbf{t}$ are at positions 3, 10, 15, 18, and 27; compare (1-2). In particular, the gap sequence is not of the form $2^\ell\mathbf{B}$ or $2^\ell\check{\mathbf{B}}$ for some $\ell \geq 0$, each of which has only three different values.

Similar considerations hold for the case $\{a_k^{\mathtt{00}}, a_k^{\mathtt{11}}\}$. More precisely, let σ_0 be the index at which w appears in $a_k^{\mathtt{11}}$ and σ_1 the index at which w appears in $a_k^{\mathtt{00}}$. Each letter occurring in $\bar{\mathbf{B}}$ corresponds to two gaps for w, as follows.

Letter in $\bar{\mathbf{B}}$	Gap 1	Gap 2
a	$\sigma_1 - \sigma_0 + 4 \cdot 2^k$	$\sigma_0 - \sigma_1 + 2 \cdot 2^k$
ā	$\sigma_1 - \sigma_0 + 2 \cdot 2^k$	$\sigma_0 - \sigma_1 + 4 \cdot 2^k$
b	$\sigma_1 - \sigma_0 + 4 \cdot 2^k$	$\sigma_0 - \sigma_1 + 4 \cdot 2^k$
c	$\sigma_1 - \sigma_0 + 2 \cdot 2^k$	$\sigma_0 - \sigma_1 + 2 \cdot 2^k$.

An example for this case is given by the word $\mathtt{00110}$, which is a factor of $a_2^{\mathtt{00}} = \mathtt{01100110}$ and of $a_2^{\mathtt{11}} = \mathtt{10011001}$. We have $\sigma_0 = 1$ and $\sigma_1 = 3$, and therefore the gaps 6, 10, 14, and 18, which appear as pairs $(18,6)$, $(10,14)$, $(18,14)$, and $(10,6)$.

3. The structure of the sequence A

In this section we investigate the infinite word $\mathbf{A}$, in particular by extending it to a word over a seven-letter alphabet. This extension allows us to better understand the structure of $\mathbf{A}$, and gives us a tool to handle the discrepancy D_N. In particular, we prove Theorem 1.2.

3.1. A is automatic. It has been known since Berstel [7] that $\mathbf{A}$ is 2-automatic. In this section we re-prove this statement using slightly different notation. Note that we give similar proofs (of Lemmas 2.2 and 2.3) in the first part of this paper. First of all, we recapture Berstel's 2-uniform morphism. Introducing an auxiliary letter $\bar{\mathsf{b}}$, we have the morphism $\overline{\varphi}$ as well as the coding π:

$$\begin{array}{lllll} \overline{\varphi}: & \mathsf{a} \mapsto \mathsf{ab}, & \mathsf{b} \mapsto \mathsf{ca}, & \bar{\mathsf{b}} \mapsto \mathsf{ac}, & \mathsf{c} \mapsto \mathsf{c}\bar{\mathsf{b}}, \\ \pi: & \mathsf{a} \mapsto \mathsf{a}, & \mathsf{b} \mapsto \mathsf{b}, & \bar{\mathsf{b}} \mapsto \mathsf{b}, & \mathsf{c} \mapsto \mathsf{c}. \end{array}$$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

We wish to prove that

$$\pi(\overline{\mathbf{A}}) = \mathbf{A}, \tag{3-1}$$

where $\overline{\mathbf{A}}$ is the fixed point of $\overline{\varphi}$ starting with $\mathtt{a}$. For this purpose, we show, by induction on $k \geq 0$, that the initial segment

$$s_k := \overline{\varphi}^k(\mathtt{abc})$$

of $\overline{\mathbf{A}}$, of length $3 \cdot 2^k$, is a concatenation of the three words $w_0 = \mathtt{abc}$, $w_1 = \mathtt{ac}$, and $w_2 = \mathtt{\bar{b}}$. We also call the words w_j 'base words' in this context, and the latter statement the 'concatenation property'. Having proved this property, we use (recall the morphism φ defined in (2-1))

$$\begin{aligned} \pi(\overline{\varphi}(w_1)) &= \mathtt{abcacb} &&= \varphi(\pi(w_1)), \\ \pi(\overline{\varphi}(w_2)) &= \mathtt{abcb} &&= \varphi(\pi(w_1)), \\ \pi(\overline{\varphi}(w_3)) &= \mathtt{ac} &&= \varphi(\pi(w_3)), \end{aligned}$$

in order to obtain

$$\varphi(\pi(s_k)) = \pi(\overline{\varphi}(s_k)) \tag{3-2}$$

for all $k \geq 0$, by concatenation. In other words, φ and $\overline{\varphi}$ act in the same way on an initial segment of $\overline{\mathbf{A}}$ of length $3 \cdot 2^k$.

We may also display the relation (3-2) graphically. Define $S = \{s_k : k \geq 0\} \subseteq \{\mathtt{a}, \mathtt{b}, \mathtt{\bar{b}}, \mathtt{c}\}^{\mathbb{N}}$ and $\Omega = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}^{\mathbb{N}}$. Then the following diagram is commutative:

$$\begin{array}{ccc} S & \xrightarrow{\pi} & \Omega \\ {\scriptstyle \overline{\varphi}}\downarrow & & \downarrow{\scriptstyle \varphi} \\ S & \xrightarrow[\pi]{} & \Omega \end{array}$$

Gluing together copies of this diagram, we obtain, for all $\ell \geq 1$,

$$\begin{aligned} \varphi^\ell(s_0) &= \varphi^\ell(\pi(s_0)) = \varphi^{\ell-1}(\pi(\overline{\varphi}(s_0))) \\ &= \varphi^{\ell-2}(\varphi(\pi(\overline{\varphi}(s_0)))) = \varphi^{\ell-2}(\pi(\overline{\varphi}^2(s_0))) = \cdots = \pi(\overline{\varphi}^\ell(s_0)). \end{aligned}$$

For each index $j \geq 0$, choose ℓ so large that $3 \cdot 2^\ell \geq j$. Then

$$\mathbf{A}_i = \varphi^\ell(s_0)|_i = \pi(\overline{\varphi}^\ell(s_0))|_i = \pi(\overline{\varphi}^\ell(s_0)|_i) = \pi(\overline{\mathbf{A}}_i)$$

for $0 \leq i < j$. Therefore the infinite word $\mathbf{A}$ is 2-automatic, being the coding under π of the 2-automatic sequence $\overline{\mathbf{A}}$, and thus we have derived (3-1) from the concatenation property.

We still have to prove that s_k is a concatenation of the base words. Clearly, this holds for $s_0 = \mathtt{abc} = w_0$. Assume that we have already established that $s_k = w_{\varepsilon_0} w_{\varepsilon_1} \cdots$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

for some $\varepsilon_j \in \{0, 1, 2\}$. We have

$$\begin{aligned}\overline{\varphi}(w_1) &= \mathtt{abcac\bar{b}} = w_0 w_1 w_2,\\ \overline{\varphi}(w_2) &= \mathtt{abc\bar{b}} = w_0 w_2,\\ \overline{\varphi}(w_3) &= \mathtt{ac} = w_1,\end{aligned}$$

and thus

$$s_{k+1} = \overline{\varphi}(s_k) = \overline{\varphi}(w_{\varepsilon_0})\overline{\varphi}(w_{\varepsilon_1})\cdots$$

is a concatenation of the w_j too. This proves (3-1).

Complementing this result, we note that Berstel [7, Corollary 7] also proved that **A** itself is not a fixed point of (the extension of) a uniform morphism.

3.2. Transforming A. We identify *circular shifts*, or *rotations*, of factors of length $L \geq 2$ appearing in the sequence **A**. Such a rotation of a word $(a_i)_{i\geq 0}$ replaces the subword $a_j\, a_{j+1} \cdots a_{j+L-2}\, a_{j+L-1}$ by $a_{j+1} \cdots a_{j+L-2}\, a_{j+L-1}\, a_j$ (rotation to the left), or $a_{j+L-1}\, a_j\, a_{j+1} \cdots a_{j+L-2}$ (rotation to the right).

Carrying out a certain number of such rotations, we see that the sequence **A** is reduced to the periodic word $(\mathtt{abc})^\omega$. Of course, this is possible for any word containing an infinite number of each of $\mathtt{a}$, $\mathtt{b}$, and $\mathtt{c}$, and it can be achieved in uncountably many ways. In our case, however, an admissible sequence of rotations can be made very explicit, by defining a new morphism φ^+. This morphism has the fixed point $\overline{\mathbf{A}}$, which maps to **A** under a coding. From this augmented sequence, we see very clearly the 'nested structure' of the above-mentioned rotations. In particular, we can find a certain *noncrossing matching*, defined in (3-8), describing the intervals that we perform rotations on, and the direction of each rotation. Moreover, in the process we learn something about the discrepancy of $\mathtt{01}$-blocks in **t**, which was defined in (1-4). Let us consider the iteration $\overline{\varphi}^2$ of Berstel's morphism:

$$\overline{\varphi}^2 : \quad \mathtt{a} \mapsto \mathtt{abca}, \quad \mathtt{b} \mapsto \mathtt{c\bar{b}ab}, \quad \mathtt{\bar{b}} \mapsto \mathtt{abc\bar{b}}, \quad \mathtt{c} \mapsto \mathtt{c\bar{b}ac}.$$

We introduce certain decorations (*connectors*) of the letters. Their meaning becomes clear in a moment. Based on the morphism $\overline{\varphi}^2$, we define the following decorated version, which is a morphism on the seven-letter alphabet

$$K = \{\mathtt{a}, \mathtt{\bar{b}}, \mathtt{\bar{b}}, \mathtt{b}, \mathtt{b}, \mathtt{c}, \mathtt{c}\}.$$

$$\varphi^+ : \begin{array}{llll} \mathtt{a} \mapsto \mathtt{abca}, & \mathtt{\bar{b}} \mapsto \mathtt{abc\bar{b}}, & \mathtt{\bar{b}} \mapsto \mathtt{abc\bar{b}}, & \\ \mathtt{b} \mapsto \mathtt{c\bar{b}ab}, & \mathtt{b} \mapsto \mathtt{c\bar{b}ab}, & \mathtt{c} \mapsto \mathtt{c\bar{b}ac}, & \mathtt{c} \mapsto \mathtt{c\bar{b}ac}. \end{array} \tag{3-3}$$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

This morphism has a unique fixed point $\mathbf{A}^+$ starting with $\mathtt{a}$. The image of $\mathbf{A}^+$ under the obvious coding γ given by

$$\gamma: \quad \begin{array}{l} \mathtt{a} \mapsto \mathtt{a}, \\ \underline{\mathtt{b}} \mapsto \mathtt{b}, \quad \underline{\mathtt{b}} \mapsto \mathtt{b}, \quad \overline{\underline{\mathtt{b}}} \mapsto \mathtt{b}, \quad \overline{\underline{\mathtt{b}}} \mapsto \mathtt{b}, \\ \underline{\mathtt{c}} \mapsto \mathtt{c}, \quad \underline{\mathtt{c}} \mapsto \mathtt{c} \end{array} \tag{3-4}$$

yields the sequence $\mathbf{A}$. Based on this, we speak of *letters of types* $\mathtt{a}$, $\mathtt{b}$, and $\mathtt{c}$, thus referring to letters from $\{\mathtt{a}\}$, $\{\overline{\underline{\mathtt{b}}}, \overline{\underline{\mathtt{b}}}, \underline{\mathtt{b}}, \underline{\mathtt{b}}\}$, and $\{\underline{\mathtt{c}}, \underline{\mathtt{c}}\}$, respectively.

From substitution (3-3), we can immediately derive the following lemma.

LEMMA 3.1. *Let* $j \geq 1$, *and* $(x, y, z) = (\mathbf{A}^+_{j-1}, \mathbf{A}^+_j, \mathbf{A}^+_{j+1})$. *Then*

$$\begin{array}{llll} y = \overline{\underline{\mathtt{b}}} & \Rightarrow & xyz = \underline{\mathtt{c}}\,\overline{\underline{\mathtt{b}}}\,\mathtt{a}, & y = \overline{\underline{\mathtt{b}}} \quad \Rightarrow \quad xyz = \underline{\mathtt{c}}\,\overline{\underline{\mathtt{b}}}\,\mathtt{a}, \\ y = \underline{\mathtt{b}} & \Rightarrow & xyz = \mathtt{a}\,\underline{\mathtt{b}}\,\underline{\mathtt{c}}, & y = \underline{\mathtt{b}} \quad \Rightarrow \quad xyz = \mathtt{a}\,\underline{\mathtt{b}}\,\underline{\mathtt{c}}. \end{array} \tag{3-5}$$

We wish to connect the 'loose ends' of the connectors; we say that two connectors at indices $i < j$ *match* if the connector at i points to the right and the connector at j points to the left. The very simple algorithm `FindMatching` joins matching connectors, beginning with shortest connections. Only pairs of *free* connectors are connected, that is, each letter may be the starting point of only one link.

```
procedure FindMatching(w):
  M←{};
  SelectedIndices←{};
  n←1;
  while n < w.length:
    for all i such that there are matching connectors at i and i+n:
      if i ∉ SelectedIndices and i+n ∉ SelectedIndices:
        Add the pair (i,i+n) to the set M;
        Add i and i+n to the set SelectedIndices;
    n←n+1;
  return M;
end.
```

Algorithm `FindMatching`: link free connectors

Note that we have to pay attention that previously selected indices are not chosen again. This explains the introduction of `SelectedIndices`. A connection between the two letters at indices i and j is just a different name for the pair (i, j). For any finite word w over the alphabet K this procedure yields a (possibly empty) set $M(w)$ of pairs (i, j) of indices.

We wish to prove that the algorithm is *monotone*.

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

LEMMA 3.2. *Let w and w' be finite words over the alphabet K, and assume that w is an initial segment of w'. Let $M(w)$ (respectively, $M(w')$) be the sets of pairs found by the* `FindMatching` *algorithm. Then*

$$M(w) \subseteq M(w').$$

PROOF. We show this by induction on the length j of w. Clearly, it holds for $j = 0$. Let us append a symbol $x \in K$ to w (at position j). Define $M_\ell(w)$ as the set of connections (a, b) for w of length strictly smaller than ℓ, found by the algorithm. Define $M_\ell(wx)$ analogously. We prove by induction on ℓ that $M_\ell(w) \subseteq M_\ell(wx)$, and that, if the inclusion is strict, we have $M_\ell(wx) = M_\ell(w) \cup \{(i,j)\}$ for some $i < j$. Suppose that this is true for some ℓ (clearly, it holds for $\ell = 0$). We distinguish between two cases. (i) If $(i,j) \notin M_\ell(wx)$ for all i, we have $M_\ell(w) = M_\ell(wx)$ by hypothesis; we add each pair (a, b) with $b < j$ having matching connectors and such that $b - a = \ell$ to the sets $M_\ell(w)$ and $M_\ell(wx)$, and possibly one more pair (i,j), for some $i < j$, to $M_\ell(wx)$. (ii) If $(i,j) \in M_\ell(wx)$ for some i, we have $\ell > j - i$ by the definition of $M_\ell(wx)$; we add the pairs (a, b), with $b < j$, having matching connectors and such that $a \neq i$ and $b - a = \ell$ to both sets $M_\ell(w)$ and $M_\ell(wx)$. There are clearly no more pairs added to $M_\ell(wx)$, since i and j are already taken; moreover, the condition that $\ell > j - i$ renders impossible the chance of another connection (i, b), where $b < j$, being added to $M_\ell(w)$. □

We extend M to a function on all (finite or infinite) words w over K, in the following obvious way: for each ℓ, form the set $\tilde{M}_\ell(w)$ of all pairs (a, b) satisfying $b - a = \ell$, having matching connectors, such that neither a nor b is a component of any $\tilde{M}_{\ell'}(w)$, where $\ell' < \ell$. Set $\tilde{M}(w) = \bigcup_{\ell \geq 1} \tilde{M}_\ell(w)$. The following lemma gives us a method to compute a matching for an infinite word by only looking at finite segments.

LEMMA 3.3. *Let w be an infinite word over K. Then*

$$\bigcup_{j \geq 0} M(w|_{[0,j)}) = \tilde{M}(w). \tag{3-6}$$

PROOF. Let $M_\ell(w)$ be the set of connections added in step ℓ of the `FindMatching` algorithm. We prove, more generally, that

$$\bigcup_{j \geq 0} M_\ell(w|_{[0,j)}) = \tilde{M}_\ell(w). \tag{3-7}$$

We prove this by induction on ℓ, and we start at connections of length $\ell = 1$. Let $(i, i+1) \in M_\ell(w|_{[0,j)})$. Then there is a pair of matching connectors at indices i and $i+1$ (where $i + 1 < j$), and therefore this pair is also contained in $\tilde{M}_1(w)$. This proves the inclusion '$\subseteq$'. On the other hand, if $(i, i+1)$ is a link connecting matching connectors in w, this link is also to be found in the sequence $w|_{[0,i+2)}$, hence the inclusion '$\supseteq$'. Assume that (3-7) holds for some $\ell \geq 1$. If the algorithm finds a pair $(i, i+\ell)$ of matching connectors in $w|_{[0,j)}$, where $(i, i+\ell) \notin M_\ell(w|_{[0,j)})$, this pair trivially also matches in the (unrestricted) word w. By hypothesis, the connectors at i and $i+\ell$ are not used by $\tilde{M}_\ell(w)$, hence the inclusion '$\subseteq$'. On the other hand, a link $(i, i+\ell)$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

of matching connectors in w that is still free in step ℓ is also free in $w|_{[0,i+\ell+1)}$ by hypothesis, which proves (3-7) and hence the lemma. □

Our algorithm avoids crossing connections: if $i < j < k < \ell$ are indices such that $(i,k) \in M(w)$ and $(j,\ell) \in M(w)$, then the connector at index j is pointing to the right, and the one at k to the left, so the shorter connection (j,k) would have been chosen earlier. This contradicts the construction rule that indices may only be chosen once.

More generally, a *noncrossing matching* for a word w over K (finite or infinite) is a set M of pairs (i,j) such that

$$\begin{aligned} &i < j && \text{for all } (i,j) \in M, \\ &w_i w_j \in \{\underline{\mathtt{b}}\mathtt{c}, \bar{\mathtt{b}}\mathtt{c}, \mathtt{c}\underline{\mathtt{b}}, \mathtt{c}\bar{\mathtt{b}}\} && \text{for all } (i,j) \in M, \\ &w_i = a && \text{for all } i \notin \bigcup M, \\ &\left.\begin{array}{ll} (i,j) = (k,\ell) & \text{or} \\ i < k < \ell < j & \text{or} \\ k < i < j < \ell & \end{array}\right\} && \text{for all } (i,j) \in M,\ (k,\ell) \in M. \end{aligned} \tag{3-8}$$

Here $\bigcup M = \{i : (i,j) \in M \text{ for some } j \text{ or } (j,i) \in M \text{ for some } j\}$.

We call a word w *closed* if there exists a noncrossing matching for w.

LEMMA 3.4. *Let w be a word over K. There is at most one noncrossing matching for w. If there exists one,* `FindMatching` *generates it by virtue of (3-6).*

PROOF. Let m be a noncrossing matching of w. Since all connectors have to connect to something and the connecting lines must not cross, we see that all pairs $(i, i+1)$ of indices where matching connectors appear have to be contained in m. It follows that $M_1(w|_{[0,j)}) \subseteq m$ for all j, and therefore $\tilde{M}_1(w) \subseteq m$ by (3-7). On the other hand, the definition of a noncrossing matching only allows matching connectors, therefore each connection $(i, i+1)$ in m is found by `FindMatching`, for $j = i+2$.

Similar reasoning applies for longer connections too. Let us assume that the set of connections of length less than ℓ coming from `FindMatching` is the same as the set of connections of length less than ℓ contained in m. Assume that i is an index such that the connectors at indices i and $i+\ell$ match, and neither i nor $i+\ell$ appears in a connection of length less than ℓ in m. Since m is a matching, the connector at index i has to be linked to a connector at an index $j > i$. Indices $j \in \{i+1, \ldots, i+\ell-1\}$ are excluded by our hypothesis, and indices $j > i+\ell$ are impossible by the noncrossing property, therefore $(i, i+\ell) \in m$. Again, other connections of length ℓ cannot appear in m, therefore `FindMatching` finds all pairs $(i, i+\ell)$ contained in m. This completes our argument by induction. Therefore, $m = \tilde{M}(w)$, and both statements of Lemma 3.4 follow. □

LEMMA 3.5. *The sequence $\mathbf{A}^+$ is closed.*

PROOF. First of all, we note that it is sufficient to prove that φ^+ maps closed words w to closed words. If this is established, we obtain, by induction, that the initial segments

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

$(\varphi^+)^k(\mathtt{a})$ of $\mathbf{A}^+$ are closed. Since noncrossing matchings are unique, the corresponding sequence $(m_k)_{k\geq 0}$ of noncrossing matchings satisfies $m_k \subseteq m_{k+1}$, and $\bigcup_{k\geq 0} m_k$ is easily seen to be the desired matching for $\mathbf{A}^+$.

We prove by induction on the length n of a closed word w that $\varphi^+(w)$ is closed. This is obvious for the closed words of length $n \leq 2$: the word $\varphi^+(\mathtt{a}) = \mathtt{a}\,\underbrace{\mathtt{b}\,\mathtt{c}}\,\mathtt{a}$ is closed, and the cases $\underbrace{\mathtt{b}\,\mathtt{c}}$, $\underbrace{\bar{\mathtt{b}}\,\mathtt{c}}$, $\underbrace{\mathtt{c}\,\mathtt{b}}$, and $\underbrace{\mathtt{c}\,\bar{\mathtt{b}}}$ are also easy. Moreover, a concatenation of two closed words is also closed: one of the matchings has to be shifted (both components of each entry have to be shifted), and we only have to form the union of the matchings.

If w is of the form $\underbrace{\bar{\mathtt{b}}\,C\,\mathtt{c}}$ for some nonempty word C over K, we obtain a noncrossing matching for C by stripping the pair $(1, n)$ from a corresponding matching for w. Therefore, C is closed. Applying φ^+, we see that

$$\varphi^+(w) = \mathtt{a}\,\underbrace{\mathtt{b}\,\mathtt{c}}\,\underbrace{\bar{\mathtt{b}}}\,\varphi^+(C)\,\underbrace{\mathtt{c}\,\bar{\mathtt{b}}}\,\mathtt{a}\,\underbrace{\mathtt{c}}. \tag{3-9}$$

This is closed by our hypothesis, since C is shorter than w. The other case $\underbrace{\mathtt{c}\,C\,\mathtt{b}}$ is analogous (note that there are no more cases by (3-5)), and the proof is complete. □

REMARK 3.6. We note that this proof can also be used to show that the substitution φ^+ *respects* noncrossing matchings, in the following sense. If m is a noncrossing matching for w, then there exists a (unique) noncrossing matching m' for $\varphi^+(w)$; the matching m can be recovered from m' by omitting certain links, and applying a renaming $(i, j) \mapsto (\mu(i), \mu(j))$ to the remaining links, where $\mu : \mathbb{N} \to \mathbb{N}$ is nonincreasing. The proof is not difficult: if this procedure works for the closed word C, we can also carry this out for $\underbrace{\bar{\mathtt{b}}\,C\,\mathtt{c}}$ by (3-9); we see that the additional link $\underbrace{\bar{\mathtt{b}}\cdots\mathtt{c}}$ is still present in $\varphi^+(\underbrace{\bar{\mathtt{b}}\,C\,\mathtt{c}})$. Also, the procedure of recovering m' from m is compatible with concatenations of closed words C and D, as a matching for $\varphi^+(CD) = \varphi^+(C)\,\varphi^+(D)$ does not connect letters in $\varphi^+(C)$ and $\varphi^+(D)$.

The construction of the matching in the proof of Lemma 3.5 also shows the following result.

COROLLARY 3.7. *Let m be the noncrossing matching for $\mathbf{A}^+$. By virtue of m, each letter of type* c *is connected to exactly one letter, which is of type* b*, and each letter of type* b *is connected to exactly one letter, which is of type* c.

Our interest in the link structure of $\mathbf{A}^+$ stems from the fact that we may transform the sequence $\mathbf{A}$ into a periodic one, using the following transparent mechanism. Let m be the noncrossing matching for $\mathbf{A}^+$, and let $((i_k, j_k))_{k\geq 0}$ be an enumeration of m such that $(j_k - i_k)_{k\geq 0}$ is nondecreasing. We define a sequence $(A^{(k)})_{k\geq 0}$ as follows.

- Set $A^{(0)} = \mathbf{A}^+$.
- Let $k \geq 0$. If $A^{(k)}_{i_k} = \underbrace{\mathtt{c}}$, we *rotate* the letters in $A^{(k)}$ with indices $i_k, i_k + 1, \ldots, j_k$ *to the right* by one place, yielding $A^{(k+1)}$. Otherwise, we necessarily have $A^{(k)}_{j_k} = \underbrace{\mathtt{c}}$ and we rotate the letters with indices $i_k, i_k + 1, \ldots, j_k - 1$ *to the left* by one place.

In more colourful language, in each step some letter of type b is moved along its connecting link and inserted just before the letter of type c it is connected to. Note

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

that due to the monotonicity requirement and the noncrossing property, the kth rotation does not change the indices at which the subsequent rotations are carried out. Therefore, the sequence $(A^{(k)})_{k\geq 0}$ is well defined. Moreover, the result does not depend on the particular nondecreasing enumeration of m for the same reasons. Since the first N indices eventually remain unchanged, the limit

$$\rho(\mathbf{A}^+) := \gamma\Big(\lim_{k\to\infty} A^{(k)}\Big)$$

exists (note that γ, defined in (3-4), replaces each letter of type x by x). The definition of $A^{(k)}$ is summarized in the `RotateAlongLinks` algorithm. As in the case of the `FindMatching` algorithm, we require a finite word w (and a finite set $m \subset \mathbb{N}^2$) as input in order to guarantee finite running time.

```
procedure RotateAlongLinks(w,m)
  if m is not a noncrossing matching for w:
    exit(Error: a noncrossing matching is required);
  Create a list m' from m, ordered by SecondComponent-FirstComponent
  for p in m':
    i←p.FirstComponent;
    j←p.SecondComponent;
    if w[i]=c:
      #Rotate right
      (w[i],…,w[j-1],w[j])←(w[j],w[i],…,w[j-1]);
    else:
      #In this case, w[j]=c. Rotate left
      (w[i],w[i+1],…,w[j-1])←(w[i+1],…,w[j-1],w[i]);
  return w;
end.
```

`RotateAlongLinks` algorithm: transform a closed word according to a noncrossing matching

By the above remarks, the words `RotateAlongLinks` $(w_k, M(w_k))$ converge to $\rho(\mathbf{A}^+)$ as $k \to \infty$, where $w = (\varphi^+)^k(\mathtt{a})$. We have the following central proposition.

PROPOSITION 3.8. *Let m be the noncrossing matching for $\mathbf{A}^+$. Then*

$$\rho(\mathbf{A}^+) = (\mathtt{abc})^\omega.$$

PROOF. Let us first note that the limit itself can be obtained in a simpler way. For any closed word C over K, (1) apply γ, (2) remove all occurrences of b, and (3) reinsert b before each c. The resulting word equals $\rho(C)$. This statement simply follows from the facts that (i) both procedures do not change the order in which the underlying letters a and c appear, that (ii) each occurrence of c in both results is preceded by b, and that (iii) in both results, b does not appear at other places. We therefore see that Proposition 3.8 is equivalent to the following. Let $\mathbf{C}$ be the sequence obtained from $\mathbf{A}^+$ by deleting

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

all decorations, and all occurrences of $\mathtt{b}$ and $\bar{\mathtt{b}}$. Then $\mathbf{C} = (\mathtt{ac})^\omega$. In other words, we only have to show that $\mathtt{a}$ and $\mathtt{c}$ occur alternatingly in $\mathbf{A}$, with the empty word or one occurrence of $\mathtt{b}$ in between. We prove a stronger statement concerning the sequence $\overline{\mathbf{A}}$, which completes the proof of Proposition 3.8.

LEMMA 3.9. *There are sequences* $(\varepsilon_k)_{k\geq 0}$ *and* $(\varepsilon'_k)_{k\geq 0}$ *in* $\{0,1\}$ *such that*

$$\overline{\mathbf{A}} = \mathtt{a}(\mathtt{bc}(\mathtt{ac})^{\varepsilon_0}\bar{\mathtt{b}}\mathtt{a}(\mathtt{ca})^{\varepsilon'_0})(\mathtt{bc}(\mathtt{ac})^{\varepsilon_1}\bar{\mathtt{b}}\mathtt{a}(\mathtt{ca})^{\varepsilon'_1})\cdots.$$

In order to prove this, we apply the second iteration $\overline{\varphi}^2$ of Berstel's morphism to one of the expressions in brackets. We use the abbreviation

$$b(\varepsilon,\varepsilon') = \mathtt{bc}(\mathtt{ac})^{\varepsilon}\bar{\mathtt{b}}\mathtt{a}(\mathtt{ca})^{\varepsilon'}.$$

Direct computation yields

$$\begin{aligned}
\overline{\varphi}^2(b(0,0)) &= \mathtt{c}\bar{\mathtt{b}}\mathtt{abc}\bar{\mathtt{b}}\mathtt{acabc}\bar{\mathtt{b}}\mathtt{abca} = \mathtt{c}\bar{\mathtt{b}}\mathtt{a}\,b(0,1)b(0,0)\mathtt{bca},\\
\overline{\varphi}^2(b(0,1)) &= \mathtt{c}\bar{\mathtt{b}}\mathtt{abc}\bar{\mathtt{b}}\mathtt{acabc}\bar{\mathtt{b}}\mathtt{abcac}\bar{\mathtt{b}}\mathtt{acabca} = \mathtt{c}\bar{\mathtt{b}}\mathtt{a}\,b(0,1)b(0,0)b(1,1)\mathtt{bca},\\
\overline{\varphi}^2(b(1,0)) &= \mathtt{c}\bar{\mathtt{b}}\mathtt{abc}\bar{\mathtt{b}}\mathtt{acabcac}\bar{\mathtt{b}}\mathtt{acabc}\bar{\mathtt{b}}\mathtt{abca} = \mathtt{c}\bar{\mathtt{b}}\mathtt{a}\,b(0,1)b(1,1)b(0,0)\mathtt{bca},\\
\overline{\varphi}^2(b(1,1)) &= \mathtt{c}\bar{\mathtt{b}}\mathtt{abc}\bar{\mathtt{b}}\mathtt{acabcac}\bar{\mathtt{b}}\mathtt{acabc}\bar{\mathtt{b}}\mathtt{abcac}\bar{\mathtt{b}}\mathtt{acabca}\\
&= \mathtt{c}\bar{\mathtt{b}}\mathtt{a}\,b(0,1)b(1,1)b(0,0)b(1,1)\mathtt{bca}.
\end{aligned}$$

Arbitrary concatenations of these expressions are again of the form $\mathtt{c}\bar{\mathtt{b}}\mathtt{a}R\,\mathtt{bca}$, where R is a concatenation of words $b(\varepsilon,\varepsilon')$. Assuming that w is of the form $\mathtt{a}\prod_{j<r} b(\varepsilon_j,\varepsilon'_j)\mathtt{bca}$, we obtain $\overline{\varphi}^2(w) = \mathtt{a}\,b(1,0)R\,\mathtt{bca}$. Since the words $(\overline{\varphi}^2)^k(\mathtt{a})$ approach a fixed point, and

$$\overline{\varphi}^4(\mathtt{a}) = \mathtt{a}\,b(1,0)b(0,1)\mathtt{bca},$$

it follows by induction that $\overline{\mathbf{A}}$ is indeed of the form stated in the lemma, and we have, in particular, proved Proposition 3.8. □

From this algorithm, we can clearly see that a given letter $\mathtt{a}$ is shifted, one place at a time, for each link that is passing over this letter. The direction in which $\mathtt{a}$ is shifted depends on whether $\underset{\llcorner}{\mathtt{c}}$ or $\underset{\lrcorner}{\mathtt{c}}$ appears in the link we are dealing with. We use considerations of this kind in the following section, together with Proposition 3.8, in order to determine the *discrepancy* of $\mathtt{01}$-occurrences in $\mathbf{t}$.

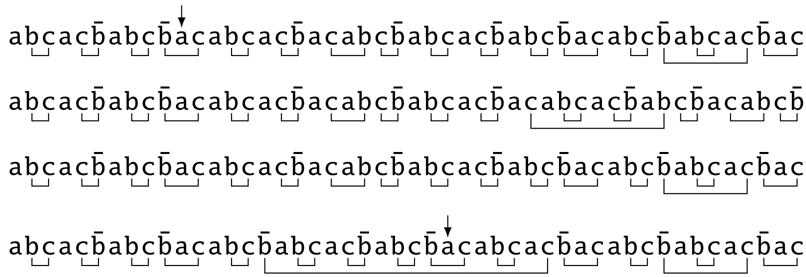
3.3. The discrepancy of 01-blocks. For an integer $j \geq 0$ let us define the *degree* of j as follows. Let m be the noncrossing matching for $\mathbf{A}^+$ and set

$$\begin{aligned}
\deg^+(j) &= \#\{(k,\ell)\in m : k<j<\ell \text{ and } \mathbf{A}^+_k = \underset{\llcorner}{\mathtt{c}}\},\\
\deg^-(j) &= \#\{(k,\ell)\in m : k<j<\ell \text{ and } \mathbf{A}^+_\ell = \underset{\lrcorner}{\mathtt{c}}\},\\
\deg(j) &= \deg^+(j) - \deg^-(j).
\end{aligned}$$

We also talk about the degree of a letter in $\mathbf{A}^+$, where the position in question will always be clear from the context.

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

We display the first 192 letters of $\mathbf{A}^+$, obtained by applying the third iteration of φ^+ to the word $\mathbf{A}^+{}_0\mathbf{A}^+{}_1\mathbf{A}^+{}_2$ = abc, and we connect associated connectors by actual lines for better readability. In position $10 = (22)_4$ in $\mathbf{A}^+$, we have a letter a of degree -1, and in position $170 = (2222)_4$, a letter a of degree -2. These positions are marked with an arrow.

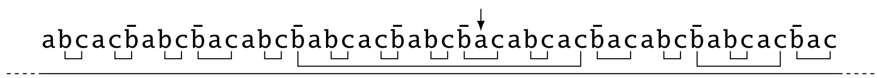


(3-10)

From this initial segment we see that the sequence $(\deg(j))_{j\ge 0}$ starts with the 48 integers

$$0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,$$
$$-1, 0, 0, 0, 0, 0, -1, -1, -1, -1, 0, 0, -1, 0, \ldots,$$

corresponding to the first line of (3-10). Applying the substitution $(\varphi^+)^2$ to b̄ac appearing in positions 169–171, we obtain the following 48 letters. The marked letter a has degree -3, and it corresponds to the position $(222222)_4$.



In general, in position $(2^{2k})_4$, a letter a of degree $-k$ appears. This can be seen by considering the images of a and c under φ^+.

By Proposition 3.8, $\deg(j)$ has the following meaning in the case where $\mathbf{A}^+{}_j$ = a. A number of letters b (of which there are $\deg^+(j)$) are transferred from the right of the letter a to the left of it; note that the letter a is shifted to the right $\deg^+(j)$ places. Analogously, $\deg^-(j)$ letters b̄ are transferred from the left of a to the right, and the letter a is shifted to the left $\deg^-(j)$ places. In total, the letter a (among other letters) is shifted by $\deg(j)$ places, and bs or b̄s are moved to account for the generated trailing space. The proposition states that the letters to the left of a's new position $j + \deg^+(j)$ are balanced: after removing decorations and replacing b̄ by b, the letters a, b, and c occur the same number of times. If $\mathbf{A}^+{}_j \in \{$c, c$\}$, similar considerations hold. The case of letters of type b is different, since a single rotation may shift such a letter to a remote place.

The transducer $\mathcal{T}_1$ displayed in Figure 1 allows us to compute the degree of an arbitrary position j: starting from the centre node, we traverse the graph, guided by the base-4 expansion $\delta_{\nu-1}\cdots\delta_0$ of j (read from left to right). Along the way, we sum up the numbers k whenever a vertex $\delta_i \mid k$ is taken. The sum over these numbers is the degree of j, multiplied by 3. The transducer $\mathcal{T}_1$ is derived directly from the decorated,

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

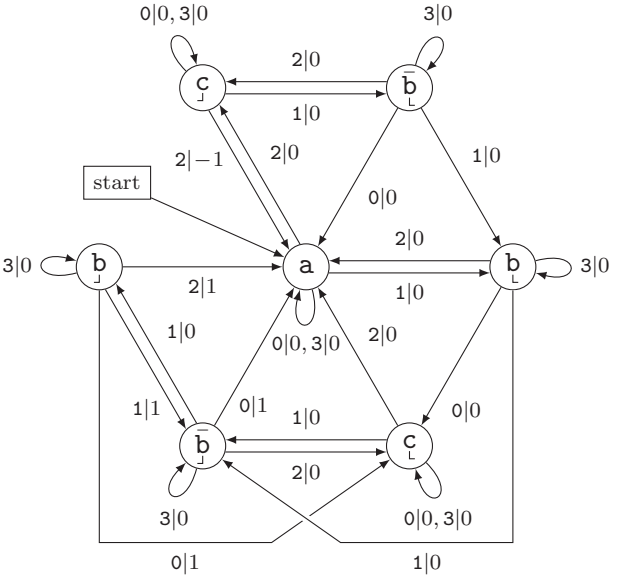


FIGURE 1. A base-4 transducer that generates the degree sequence.

4-uniform morphism φ^+ given in (3-3). Note that a change of degree takes place whenever new letters are inserted, by virtue of the morphism φ^+, into the range of already existing links, which happens for $\underline{\mathtt{b}}$ and $\underline{\bar{\mathtt{b}}}$; or if a new link together with a letter $\mathtt{a}$ in its range is created, which happens for $\underline{\mathtt{c}}$.

We now apply Proposition 3.8 to the discrepancy D_N of occurrences of $\mathtt{01}$ in $\mathbf{t}$.

PROPOSITION 3.10. *Let $j \in \mathbb{N}$ and set $d = \deg(j)$. Then*

$$\begin{aligned}
D_{4j} &= d/3, && \text{if } \mathbf{A}^+_j = \mathtt{a},\\
D_{4j} &= d/3 + 1/3, && \text{if } \mathbf{A}^+_j = \underline{\bar{\mathtt{b}}},\\
D_{4j} &= d/3, && \text{if } \mathbf{A}^+_j = \underline{\bar{\mathtt{b}}},\\
D_{4j+2} &= d/3 + 1/3, && \text{if } \mathbf{A}^+_j = \underline{\mathtt{b}},\\
D_{4j+2} &= d/3, && \text{if } \mathbf{A}^+_j = \underline{\bar{\mathtt{b}}},\\
D_{4j+2} &= d/3 - 1/3, && \text{if } \mathbf{A}^+_j = \underline{\mathtt{c}},\\
D_{4j+2} &= d/3, && \text{if } \mathbf{A}^+_j = \underline{\mathtt{c}}.
\end{aligned} \tag{3-11}$$

In each of these cases, the subscript of D is the position in $\mathbf{t}$ that corresponds to the jth letter in $\mathbf{A}$ via (2-3).

PROOF. Choose $\varepsilon \in \{0, 1, 2\}$ and $n \in \mathbb{N}$ such that $j = 3n + \varepsilon$. Let us consider each of the seven cases corresponding to letters from K.

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

First case. Assume that $\mathbf{A}^+_j = \mathtt{a}$. By the `RotateAlongLinks` algorithm and Proposition 3.8, a total of d letters of type $\mathtt{b}$ have to be shifted from the right of our $\mathtt{a}$ in question to the left (if $d > 0$), or the other way round (if $d < 0$). After this procedure, the numbers of letters of types $\mathtt{a}$, $\mathtt{b}$, and $\mathtt{c}$ to the left are equal. It follows that $\varepsilon \equiv -d \bmod 3$; moreover, $m = n + (\varepsilon + d)/3$ is the number of letters $\mathtt{a}$ (and also the number of letters of type $\mathtt{c}$) strictly to the left of j. The number of letters of type $\mathtt{b}$ to the left of j is $m' = n + (\varepsilon - 2d)/3$. Symbols of type $\mathtt{a}$ contribute two blocks $\mathtt{01}$ and correspond to a factor of length 6 in $\mathbf{t}$, by (2-2); letters of type $\mathtt{b}$ contribute one block and correspond to a factor of length 4; letters of type $\mathtt{c}$ contribute one block and correspond to a factor of length 2. It follows that below position

$$N = (6+2)\left(n + \frac{\varepsilon + d}{3}\right) + 4\left(n + \frac{\varepsilon - 2d}{3}\right) = 12n + 4\,\varepsilon = 4j$$

we find

$$(2+1)\left(n + \frac{\varepsilon + d}{3}\right) + \left(n + \frac{\varepsilon - 2d}{3}\right) = 4n + 4\,\varepsilon/3 + d/3$$

$\mathtt{01}$-blocks. This proves the case $\mathbf{A}^+_j = a$.

Second case. If $\mathbf{A}^+_j = \underline{\bar{\mathtt{b}}}$, we note that necessarily $\mathbf{A}^+_{j+1} = \mathtt{a}$, by Lemma 3.1. We apply the first case in position $j+1$, which has degree d. Noting that a letter of type $\mathtt{b}$ in $\mathbf{A}^+$ corresponds to $\mathtt{0110}$ in Thue–Morse, we obtain $D_{4j} = D_{4j+4} + 1/3 = d/3 + 1/3$, where $4j$ (respectively, $4j+2$) corresponds to the jth (respectively, $(j+1)$th) position in $\mathbf{A}^+$.

Third case. Assume that $\mathbf{A}^+_j = \underline{\bar{\mathtt{b}}}$. In this case, the letter at $j+1$ is $\mathtt{a}$ by Lemma 3.1 and $j+1$ has degree $d-1$. It follows that $D_{4j} = D_{4j+4} + 1/3 = d/3 - 1/3 + 1/3 = d/3$.

Fourth case. If $\mathbf{A}^+_j = \underline{\mathtt{b}}$, we note that necessarily $\mathbf{A}^+_{j-1} = \mathtt{a}$; we apply the first case in position $j-1$, which has degree $d+1$. Since $\mathtt{a}$ corresponds to $\mathtt{011010}$ in Thue–Morse, we have $D_{4j+2} = D_{4j-4} = d/3 + 1/3$, where $4j-4$ (respectively, $4j+2$) corresponds to the $(j-1)$th (respectively, jth) positions in $\mathbf{A}^+$.

Fifth case. Assume that $\mathbf{A}^+_j = \underline{\mathtt{b}}$. Then $\mathbf{A}^+_{j-1} = \mathtt{a}$, and $j-1$ has degree d. Analogously to the fourth case, we obtain $D_{4j+2} = D_{4j-4} = d/3$.

Sixth case. If $\mathbf{A}^+_j = \underline{\mathtt{c}}$, this letter is connected to a letter of type $\mathtt{b}$ to the left, which stays on the left of $\underline{\mathtt{c}}$ after applying the rotations. Therefore, the number of letters of type $\mathtt{b}$ to the left is changed by d, and the numbers of letters of types $\mathtt{a}$ or $\mathtt{c}$ to the left stay the same. Similarly to the first case, it follows that $\varepsilon \equiv 2 - d \bmod 3$. The numbers of letters, of type $\mathtt{a}$, $\mathtt{b}$, and $\mathtt{c}$, to the left of j, are therefore $m = n + (\varepsilon + d + 1)/3$, $m - d$, and $m - 1$, respectively. It follows that, below position

$$\begin{aligned} N &= 6\left(n + \frac{\varepsilon + d + 1}{3}\right) + 4\left(n + \frac{\varepsilon - 2d + 1}{3}\right) + 2\left(n + \frac{\varepsilon + d - 2}{3}\right) \\ &= 12n + 4\,\varepsilon + 2 = 4j + 2, \end{aligned}$$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

there are

$$2\left(n+\frac{\varepsilon+d+1}{3}\right)+\left(n+\frac{\varepsilon-2d+1}{3}\right)+\left(n+\frac{\varepsilon+d-2}{3}\right)=4n+\frac{4\varepsilon}{3}+\frac{d}{3}+\frac{1}{3}$$

$\mathtt{01}$-blocks.

Seventh case. If $\mathbf{A}^+_j = \mathtt{c}_{\llcorner}$, a letter of type $\mathtt{b}$ takes its place after one rotation. In this case, we have $\varepsilon \equiv 1-d \bmod 3$; the numbers of letters to the left of j, of types $\mathtt{a}$, $\mathtt{b}$, and $\mathtt{c}$, are therefore $m = n+(\varepsilon+d+2)/3$, $m-d-1$, and $m-1$, respectively. Therefore, below position

$$\begin{aligned} N &= 6\left(n+\frac{\varepsilon+d+2}{3}\right)+4\left(n+\frac{\varepsilon-2d-1}{3}\right)+2\left(n+\frac{\varepsilon+d-1}{3}\right)\\ &= 12n+4\varepsilon+2 = 4j+2, \end{aligned}$$

there are

$$2\left(n+\frac{\varepsilon+d+2}{3}\right)+\left(n+\frac{\varepsilon-2d-1}{3}\right)+\left(n+\frac{\varepsilon+d-1}{3}\right)=4n+\frac{4\varepsilon}{3}+\frac{d}{3}+\frac{2}{3}$$

$\mathtt{01}$-blocks, which proves the last case. □

Since $\deg(j)$ is easy to obtain, Proposition 3.10 gives us a simple method to compute the discrepancy D_N for any given N.

PROPOSITION 3.11. *Let $N \geq 0$ be an integer and $j = \lfloor N/4 \rfloor$.*

(1) *If $\mathbf{A}^+_j \in \{\mathtt{a}, \mathtt{b}_{\lrcorner}, \mathtt{b}_{\llcorner}\}$, choose $\delta = D_{4j}/3 = \deg(j)/3 + \varepsilon$, where $\varepsilon \in \{0, 1/3\}$ is given by the first block of (3-11). Then*

$$(D_{4j}, D_{4j+1}, D_{4j+2}, D_{4j+3}) = (\delta, \delta+2/3, \delta+1/3, \delta). \tag{3-12}$$

(2) *If $\mathbf{A}^+_j \in \{\mathtt{b}, \mathtt{b}_{\lrcorner}, \mathtt{c}_{\lrcorner}, \mathtt{c}_{\llcorner}\}$, choose $\delta = D_{4j+2}/3 = \deg(j)/3 + \varepsilon$, where the variable $\varepsilon \in \{-1/3, 0, 1/3\}$ is given by the second block of (3-11). Then*

$$(D_{4j}, D_{4j+1}, D_{4j+2}, D_{4j+3}) = (\delta+2/3, \delta+1/3, \delta, \delta+2/3). \tag{3-13}$$

The scaled sequence of discrepancies (multiplied by 3) therefore begins with the 48 integers

$$\begin{array}{llllll} 0,2,1,0, & 2,1,0,2, & 1,0,-1,1, & 0,2,1,0, & 2,1,0,2, & 1,3,2,1,\\ 0,2,1,0, & 2,1,0,2, & 1,0,-1,1, & 0,2,1,0, & -1,1,0,-1, & 1,0,-1,1. \end{array}$$

The partition into segments of length 4 is for better readability. Each segment corresponds to one symbol in $\mathbf{A}^+$.

PROOF OF PROPOSITION 3.11. For the first sentence of each of the two cases, there is nothing to show, by Proposition 3.10. Let us begin with the first case. By the proposition, the position $4j$ in the Thue–Morse sequence corresponds to a letter $\mathtt{a}$ or $\mathtt{b}$ in $\mathbf{A}$ (in position j), and by (2-2) we have $(\mathbf{t}_{4j}, \mathbf{t}_{4j+1}, \mathbf{t}_{4j+2}, \mathbf{t}_{4j+3}) = (\mathtt{0110})$. Therefore, (3-12) follows. Concerning the second case, Proposition 3.10 gives us an expression

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

for D_{4j+2} in terms of $\deg(j)$, and the position $4j+2$ corresponds to the index j in $\mathbf{A}$. By (2-2), we have $(\mathbf{t}_{4j+2}, \mathbf{t}_{4j+3}) = (\mathtt{0}, 1)$. Therefore,

$$(D_{4j+2}, D_{4j+3}) = (\delta, \delta + 2/3).$$

In order to compute D_{4j} and D_{4j+1} in this case, we note that $\underline{\mathtt{b}}$ and $\underline{\mathtt{b}}$ are always preceded by $\mathtt{a}$ or a letter of type $\mathtt{b}$ (as we noted in the proof of Proposition 3.10), and $\underline{\mathtt{c}}$ and $\underline{\mathtt{c}}$ are always preceded by a letter equal to $\mathtt{a}$ or of type $\mathtt{b}$, since $\mathbf{A}$ is squarefree. It follows that the letter at index $j-1$ is of type $\mathtt{a}$ or $\mathtt{b}$, and therefore $(\mathbf{t}_{4j}, \mathbf{t}_{4j+1}) = (\mathtt{10})$. Consequently, we have

$$(D_{4j}, D_{4j+1}) = (\delta + 2/3, \delta + 1/3),$$

and (3-13) follows. □

3.4. Proof of Theorem 1.2. We may now show that the sequence $(D_N)_{N\geq 0}$ of discrepancies is given by a base-2 transducer. The transducer in Figure 1 may be described by eight 7×7 matrices $A^{(\ell)}, W^{(\ell)}$, for $0 \leq \ell < 4$, where rows and columns are indexed by the letters of K, in the order $(\mathtt{a}\underline{\mathtt{b}}, \bar{\mathtt{b}}, \underline{\bar{\mathtt{b}}}, \underline{\mathtt{b}}, \underline{\mathtt{c}}, \underline{\mathtt{c}})$.

The entry $A^{(\ell)}_{i,j}$ equals 1 if there is an arrow with first component equal to ℓ from the jth node to the ith node in Figure 1, and it is zero otherwise. The matrices $A^{(\ell)}$ are permutation matrices. The entry $W^{(\ell)}_{i,j}$ is the second component of the arrow from j to i with first component ℓ, if there is one, and equal to zero otherwise.

The final modification given by (3-12) and (3-13) is dealt with by four more matrices $Z^{(\ell)}$, where $0 \leq \ell < 4$. The first three columns of these matrices are given by (3-12), as follows. Define the quadruple $(q_0, q_1, q_2, q_3) = (0, 2/3, 1/3, 0)$ (containing the shifts in (3-12)), and the triple $(r_1, r_2, r_3) = (0, 1/3, 0)$ (taking care of the shifts present in the first block of (3-11)). Let $1 \leq j \leq 3$ (corresponding to the letter at which an arrow starts), and $0 \leq \ell < 4$ (a base-4 digit, the first component of the label of the arrow). There is a unique $i \in \{1, \dots, 7\}$ such that $A^{(\ell)}_{i,j} = 1$, and we set $Z^{(\ell)}_{i,j} = q_\ell + r_j$, and $Z^{(\ell)}_{i',j} = 0$ for $i' \neq i$. The remaining four columns are filled with the help of (3-13) as follows. Define $(\tilde{q}_0, \tilde{q}_1, \tilde{q}_2, \tilde{q}_3) = (2/3, 1/3, 0, 2/3)$ and $(r_4, r_5, r_6, r_7) = (1/3, 0, -1/3, 0)$. Let $4 \leq j \leq 7$ and $0 \leq \ell < 4$. There is a unique $i \in \{1, \dots, 7\}$ such that $A^{(\ell)}_{i,j} = 1$, and we set $Z^{(\ell)}_{i,j} = \tilde{q}_\ell + r_j$, and $Z^{(\ell)}_{i',j} = 0$ for $i' \neq i$.

In order to generate the discrepancy, we blow up the transducer by a factor of 28, in order to keep track of the arrow that led to the current node (that is, we need to save the previously read digit $\ell' \in \{0, 1, 2, 3\}$ and the node in $\mathcal{T}_1$ that was last visited).

In each step, the contribution of $Z^{(\ell')}$ is cancelled out, and the contributions of $A^{(\ell')}$ and $Z^{(\ell)}$ are added (where ℓ is the currently read digit). More precisely, let (i, ℓ', j), for $1 \leq i, j \leq 7$ and $0 \leq \ell' < 4$, be the 196 nodes of our new transducer $\mathcal{T}_2$. There is an arrow from (j, ℓ', k) to (i, ℓ, j') if and only if $j = j'$ and $A^{(\ell)}_{j,i} = 1$, that is, if there is an arrow from j to i in $\mathcal{T}_1$ whose label has ℓ as its first component. We may now define the weight of an arrow $(j, \ell', k) \to (i, \ell, j)$ as

$$Z^{(\ell)}_{i,j} - Z^{(\ell')}_{j,k} + W^{(\ell')}_{j,k}.$$

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

The initial node is $(1,0,1)$, which corresponds to the fact that leading zeros do not make a difference. Let us illustrate, by a short but representative example, the easy proof that the transducer $\mathcal{T}_2$ generates the discrepancy sequence. We wish to compute the discrepancy $D_{41} = D_{(221)_4}$. The corresponding path in $\mathcal{T}_2$ is given by

$$(1,0,1) \longrightarrow (5,2,1) \longrightarrow (1,2,5) \longrightarrow (4,1,1).$$

Note that the first and third components correspond to letters in K, that is, to nodes in $\mathcal{T}_1$, via $1 \rightleftharpoons \mathtt{a}$, $4 \rightleftharpoons \mathtt{b}$, and $5 \rightleftharpoons \mathtt{c}$. The sum of the weights simplifies, due to a telescoping sum and since $W^{(0)}_{1,1} = Z^{(0)}_{1,1} = 0$, to

$$W^{(2)}_{5,1} + W^{(2)}_{1,5} + Z^{(1)}_{4,1}.$$

The first two summands sum to $\deg((22)_4) = -1/3$ by the construction of our transducer, while the last summand consists of two parts: the shift in the first line of the first block of (3-11) (which is 0), and the shift in the second component of (3-12) (which is $2/3$). Summing up, we obtain $D_{41} = 1/3$. It is clear that the proof of the general case is not more complicated this example.

Since the integers 2 and 4 are multiplicatively dependent, in symbols, $2^m = 4^n$ for $(m,n) = (2,1)$, the sequence D is also generated by a base-2 transducer. In order to carry out this reduction to base 2, the four arrows starting from a given node in our base-4 transducer have to be replaced by a complete binary tree of depth 2, where two auxiliary nodes have to be inserted. This completes the proof of the proposition and thus the proof of the first part of Theorem 1.2.

The output sum of a base-q transducer is clearly bounded by a constant times the length of the base-q expansion we feed into the transducer. This immediately yields $D_N \ll \log N$.

We easily see from Figure 1 that the integers

$$(2^{2k})_4 = 2\,\frac{16^k - 1}{3} \quad \text{and} \quad ((110)^k)_4 = 20\,\frac{64^k - 1}{63}$$

have degrees $-k$ and k, respectively, for $k \geq 1$, and that the letter $\mathtt{a}$ is attained at these positions. Therefore, Proposition 3.10 implies

$$D_{8(16^k-1)/3} = -k/3 \quad \text{and} \quad D_{80(64^k-1)/63} = k/3$$

for $k \geq 1$, and clearly $D_0 = 0$. In particular, $\{D_N : N \geq 0\} = (1/3)\mathbb{Z}$, which finishes the proof of Theorem 1.2. □

By considering the path given by $n' = (2^{2k-1})_4$ instead, we end up in the node $\mathtt{c}$, and the position n' has degree $-k+1$. Proposition 3.11 implies $D_n = -k/3$, where $n = 4n' + 2 = ((10)^{4k})_2$. This was observed by Jeffrey Shallit (private communication, 2021), but such an unboundedness result does not seem to be stated in the literature.

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press

Acknowledgements

The author thanks Jeff Shallit for sharing with him the research question treated in this paper, for constant interest in his research, and for quick and informative answers to e-mails. The question was presented to the author during the workshop 'Numeration and Substitution' at the ESI Vienna (Austria) in July 2019. We express our thanks to the ESI for providing optimal working conditions at the workshop, including offices for participants and blackboards in unconventional places. Finally, we thank Michel Rigo for fruitful discussions on the topic and Clemens Müllner for pointing out that the case of arbitrary factors can be reduced to the case `01`.

References

[1] J.-P. Allouche, F. M. Dekking and M. Queffélec, 'Hidden automatic sequences', *Combinatorial Theory* **1** (2021): 20.
[2] J.-P. Allouche and J. Shallit, 'The ring of k-regular sequences', *Theoret. Comput. Sci.* **98**(2) (1992), 163–197.
[3] J.-P. Allouche and J. Shallit, 'The ubiquitous Prouhet–Thue–Morse sequence', in: *Sequences and Their Applications (Proceedings of SETA '98, Singapore, 1998)*, Discrete Mathematics and Theoretical Computer Science (eds. C. Ding, T. Helleseth and H. Niederreiter) (Springer, London, 1999), 1–16.
[4] J.-P. Allouche and J. Shallit, *Automatic Sequences: Theory, Applications, Generalizations* (Cambridge University Press, Cambridge, 2003).
[5] J.-P. Allouche and J. Shallit, 'The ring of k-regular sequences. II', *Theoret. Comput. Sci.* **307**(1) (2003), 3–29.
[6] J.-P. Allouche and J. Shallit, 'Automatic sequences are also non-uniformly morphic', in *Discrete Mathematics and Applications* (eds. A. M. Raigorodskii and M. Th. Rassias) Springer Optimization and Its Applications, 165 (Springer, Cham, 2020), 1–6.
[7] J. Berstel, 'Sur la construction de mots sans carré', in: *Séminaire de théorie des nombres (1978–1979)* (CNRS, Talence, 1979), Exp. no. 18, 15 pages.
[8] J. Berstel, *Axel Thue's Papers on Repetitions in Words: A Translation*, Publications du Laboratoire de Combinatoire et d'Informatique Mathématique, 20 (Université du Québec à Montréal, Montreal, 1995).
[9] V. Berthé and P. C. Bernales, 'Balancedness and coboundaries in symbolic systems', *Theoret. Comput. Sci.* **777** (2019), 93–110.
[10] F. Blanchet-Sadri, J. D. Currie, N. Rampersad and N. Fox, 'Abelian complexity of fixed point of morphism $0 \mapsto 012, 1 \mapsto 02, 2 \mapsto 1$', *Integers* **14** (2014), A11, 17 pages.
[11] S. J. Brams and A. D. Taylor, *The Win-Win Solution: Guaranteeing Fair Shares to Everybody* (W. W. Norton, New York, 1999).
[12] J. L. Brooks *et al., The Simpsons Movie* (20th Century Fox Home Entertainment, Beverly Hills, CA, 2007).
[13] S. Brown, N. Rampersad, J. Shallit and T. Vasiga, 'Squares and overlaps in the Thue–Morse sequence and some variants', *RAIRO Theor. Inform. Appl.* **40** (2006), 473–484.
[14] F. M. Dekking, 'Morphisms, symbolic sequences, and their standard forms', *J. Integer Seq.* **19**(1) (2016), 16.1.1, 8 pages.
[15] H. Delange, 'Sur les fonctions q-additives ou q-multiplicatives', *Acta Arith.* **21** (1972), 285–298.
[16] F. Durand, 'A characterization of substitutive sequences using return words', *Discrete Math.* **179**(1–3) (1998), 89–101.
[17] P. Erdős, C. Mauduit and A. Sárközy, 'On arithmetic properties of integers with missing digits. I. Distribution in residue classes', *J. Number Theory* **70**(2) (1998), 99–120.

[18] C. Heuberger, S. Kropf and H. Prodinger, 'Output sum of transducers: limiting distribution and periodic fluctuation', *Electron. J. Combin.* **22**(2) (2015), 2.19, 53 pages.
[19] S. Istrail, 'On irreductible languages and nonrational numbers', *Bull. Math. Soc. Sci. Math. Roumanie (N.S.)* **21**(69) (1977), 301–308.
[20] J. Justin and L. Vuillon, 'Return words in sturmian and episturmian words', *RAIRO Theor. Inform. Appl.* **34**(5) (2000), 343–356.
[21] M. Lejeune, J. Leroy and M. Rigo, 'Computing the k-binomial complexity of the Thue–Morse word', *J. Combin. Theory Ser. A* **176** (2020), 105284.
[22] M. Lothaire, *Algebraic Combinatorics on Words*, Encyclopedia of Mathematics and Its Applications, 90, A collective work by J. Berstel *et al.* with a preface by J. Berstel and D. Perrin (Cambridge University Press, Cambridge, 2002).
[23] C. Mauduit, 'Multiplicative properties of the Thue–Morse sequence', *Period. Math. Hungar.* **43**(1–2) (2001), 137–153.
[24] M. Rao, M. Rigo and P. Salimov, 'Avoiding 2-binomial squares and cubes', *Theoret. Comput. Sci.* **572** (2015), 83–91.
[25] M. Rigo and P. Salimov, 'Another generalization of abelian equivalence: binomial complexity of infinite words', *Theoret. Comput. Sci.* **601** (2015), 47–57.
[26] N. J. A. Sloane and The OEIS Foundation Inc., *The On-Line Encyclopedia of Integer Sequences*, 2021.
[27] A. Thue, 'Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen', *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67, reprinted in *Selected Mathematical Papers of Axel Thue* (T. Nagell, editor) (Universitetsforlaget, Oslo, 1977), 413–478.

LUKAS SPIEGELHOFER, Mathematics, Montanuniversität Leoben,
Leoben, Austria
e-mail: lukas.spiegelhofer@unileoben.ac.at

https://doi.org/10.1017/S1446788721000380 Published online by Cambridge University Press