On the number of frames in binary words

Tero Harju[a], Tomi Kärki[a,1,*]

[a]*Department of Mathematics and Turku Centre for Computer Science, University of Turku, 20014 Turku, Finland*

Abstract

A frame is a square uu, where u is an unbordered word. Let $F(n)$ denote the maximum number of distinct frames in a binary word of length n. We count this number for small values of n and show that $F(n)$ is at most $\lfloor n/2 \rfloor + 8$ for all n and greater than $7n/30 - \epsilon$ for any positive ϵ and infinitely many n. We also show that Fibonacci words, which are known to contain plenty of distinct squares, have only a few frames. Moreover, by modifying the Thue-Morse word, we prove that the minimum number of occurrences of frames in a word of length n is $\lceil n/2 \rceil - 2$.

Keywords: Frame, Square, Unbordered word, Fibonacci word, Thue-Morse word
2000 MSC: 68R15

1. Introduction

Since the seminal papers of Thue [1, 2] repetitions have been one of the main subjects in combinatorics on words. Here we confine ourselves to studying squares, i.e., repetitions of the form $uu = u^2$, where u is a nonempty word. Both the number of distinct squares and the number of occurrences of squares (repeated squares) is considered. Let us first recall some earlier results.

Let $D(n)$ denote the maximum number of distinct squares in a word of length n. Fraenkel and Simpson proved in [3] that $D(n) < 2n$ for all $n > 0$. Moreover, they showed that the maximum number $P(n)$ of distinct

*Corresponding author
Email addresses: harju@utu.fi (Tero Harju), topeka@utu.fi (Tomi Kärki)
[1]Present address: Department of Teacher Education, University of Turku, PO Box 175, 26101 Rauma, Finland

primitively rooted squares in a word of length n satisfies $P(n) \geq n - o(n)$ for infinitely many n. A primitively rooted square is a word u^2, where u is primitive. A short proof for the upper bound $D(n) < 2n$ was given by Ilie [4], who also obtained a better upper bound $2n - \Theta(\log n)$ in [5]. However, based on the numerical evidence [3], the conjectured bound is n.

The minimum number of distinct squares in a binary word was consider in another paper by Fraenkel and Simpson [6]. Let $g(k)$ denote the length of a longest binary word containing at most k distinct squares. How does the sequence $\{g(k)\}$ behave? It is easy to compute the first values: $g(0) = 3$, $g(1) = 7$ and $g(2) = 18$. Fraenkel and Simpson proved that $g(3) = \infty$ by constructing an infinite word containing only squares 00, 11 and 0101; for easier proofs, see [7, 8].

Let $r(n)$ denote the minimum number of occurrences of squares in a word of length n. Kucherov, Ochem and Rao showed that $r(n)/n$ converges to a constant $0.55080\cdots$ [9]. Note that the maximum number of occurrences of squares is obtained by the word 0^n, where every even length factor is a square. This gives $n^2/4$ and $(n^2-1)/4$ repeated squares for even and odd n, respectively. Moreover, Crochemore showed that a binary word can contain $\Theta(n \log n)$ occurrences of primitively rooted squares [10].

In this paper we consider variants of the above problems by estimating the number of *frames*, i.e., squares u^2, where u is an unbordered word. Unbordered words and factors of words play an important role in combinatorics on words, for example, in connection with periodicity, coding properties of sets of words and unavoidability; see, e.g., [11, 12, 13, 14, 15, 16]. Note that the minimum number of distinct frames in a binary word of length at least 19 is three. This follows directly from the infinite word containing only squares 00, 11 and 0101 constructed by Fraenkel and Simpson [6]. Moreover, the maximum number of occurrences of frames is again given by the word 0^n, which contains $n-1$ frames 00. However, the questions of finding the maximum number of distinct frames and the minimum number of occurrences of frames for words of given length are not so straightforward.

First, we consider an interesting example. Despite the fact that a Fibonacci word of length F_n has around $0.7639F_n$ distinct square factors [17], it turns out in Section 3 that Fibonacci words contain only a few frames. In Section 4 we consider an upper bound for the maximum number of distinct frames $F(n)$ in a binary word of length n. We prove that $F(n)$ is at most $\lfloor n/2 \rfloor + 8$. Moreover, using prefixes of the Thue-Morse word we construct in Section 5 arbitrarily long words attesting to $F(n) > \frac{7}{30}n - \epsilon$ for any posi-

tive ϵ and infinitely many n. Finally, in Section 6 we prove that the minimum number of occurrences of frames in a word of length n is exactly $\lceil n/2 \rceil - 2$ for $n \geq 3$.

2. Frames

We consider here binary words $w \in \{0,1\}^*$. A *frame* is a square uu, where u is an unbordered word. A word w is *unbordered* if $w = vu = u'v$ for a nonempty v implies $v = w$. Otherwise, the word w is called *bordered* and $v \neq w$ is a *border* of w. We observe that if w is bordered then it has a border v of length $|v| \leq |w|/2$. For example, a word 00010001 is a frame, since 0001 is unbordered. The set

$$S = \{00, 11, 0101, 1010, 001001, 110110, 011011, 100100\}$$

consists of all *short frames*, frames of length at most six. For a word w,

- $F(w)$ is the number of different frames in w. Also, let
$$F(n) = \max\{F(w) \mid |w| = n\}.$$
- $M(w)$ denotes the total number of occurrences of frames in w. Also, let
$$M(n) = \min\{M(w) \mid |w| = n\}.$$
- $S(w)$ is the number of occurrences of short frames from the set S. Also, let
$$S(n) = \min\{S(w) \mid |w| = n\}.$$

For instance, let $w = 001100110$. Now $F(w) = 3$ and $M(w) = 5$ since w contains the frames 00 (twice), 11 (twice), 00110011 (once). Also, $S(w) = 4$ since 00 and 11 both occur twice and 00110011 is not in S. For the length $n = 18$, one has a unique word w modulo complementing, i.e., interchanging 0 and 1 such that $F(w) = F(18)$. This is the word 011011001011001001 having nine different frames uu, where

$$u \in \{0, 1, 01, 011, 110, 001, 100, 101100, 110010\}.$$

In Table 1 we have listed the numbers $F(n)$ and $S(n)$ for small values of n.

n	6	7	8	9	10	11	12	13	14	15	16
$F(n)$	3	3	4	4	5	6	6	6	7	8	8
$S(n)$	1	2	2	3	3	4	4	5	5	6	6
n	17	18	19	20	21	22	23	24	25	26	27
$F(n)$	8	9	9	9	10	10	10	10	11	11	12
$S(n)$	7	7	8	8	9	9	10	10	11	11	12

Table 1: The maximum number of frames and the minimum number of short frames.

3. The case of Fibonacci

Let F_n be the nth Fibonacci number, i.e., the length of the nth Fibonacci word f_n. Thus $f_n = f_{n-1}f_{n-2}$ for $n \geq 2$ with $f_0 = 1$ and $f_1 = 0$. It was shown by Fraenkel and Simpson [17] that the nth Fibonacci word has $2(F_{n-2} - 1)$ distinct square factors. Asymptotically this is around $0.7639F_n$. Hence Fibonacci words have a wealth of distinct squares. However, we show that f_n has only a few frames.

A factorization $w = uv$ such that u and v are nonempty and $|u| = p$ is called *critical* if the local period at point p is equal to the global period, i.e., the minimal period of w. Then p is called a *critical point*; see [18, Section 8.2]. For the proof of the following lemma, see [19].

Lemma 1. *Let $w = uv$ be a word such that its critical point is positioned at $|u|$. Then the conjugate vu is unbordered.*

Currie and Saari have proven the following result concerning unbordered factors of Fibonacci words [20, Lemma 7]. Actually, their lemma and the word t_m is defined for all Sturmian sequences, but in the case of Fibonacci words we have $t_m = f_{m+1}$. Each primitive binary word w of length $|w| \geq 2$ has two Lyndon conjugates. These are the conjugates of w that are minimal with respect to the lexicographic orders induced by the order $0 < 1$ and its dual order $1 < 0$. It is well known that the Fibonacci words are primitive.

Lemma 2. *A word w with $|w| \geq 2$ is an unbordered factor of a Fibonacci word if and only if w is one of the two Lyndon words that are conjugates of a word $t_m = f_{m+1}$ for some $m \geq 1$.*

The next lemma describes exactly the unbordered conjugates of the Fibonacci words. In the below we adopt the notation

$$w^\bullet = v \quad \text{for } w = va \quad \text{where } a \in \{0, 1\}.$$

Lemma 3. *Each Fibonacci word f_n for $n \geq 2$ has exactly two unbordered conjugates. These are (1) $\hat{f}_n = af_{n-2}f_{n-1}^{\bullet}$, where $f_{n-1} = f_{n-1}^{\bullet}a$, and (2) $\tilde{f}_n = af_n^{\bullet}$, where $f_n = f_n^{\bullet}a$. Here $a \in \{0, 1\}$.*

Proof. Using the above notation, $f_n = f_{n-1}f_{n-2} = f_{n-1}^{\bullet}af_{n-2}$. It was shown in [21] that f_n has a unique critical point positioned after $f_{n-1}^{\bullet}$ and in this case the corresponding conjugate $af_{n-2}f_{n-1}^{\bullet}$ is unbordered by Lemma 1.

Consider then the second case. Now $f_{n+1} = f_nf_{n-1} = f_n^{\bullet}af_{n-1}$ and, by [21], the unique critical point of f_{n+1} follows the prefix $f_n^{\bullet}$, and hence the conjugate $af_{n-1}f_n^{\bullet}$ is unbordered. Here $f_n^{\bullet} = f_{n-1}v$ for a word v of length $|f_{n-2}|-1$. Thus if $af_n^{\bullet} = af_{n-1}v$ is bordered, let u be its shortest border. Then $|u| \leq |f_n|/2$. However, since $af_{n-1}f_n^{\bullet} = af_{n-1}f_{n-1}v$ is unbordered, we have $|u| > |af_{n-1}| > |f_n|/2$. This is a contradiction. Hence $af_n^{\bullet}$ is unbordered.

Moreover, by Lemma 2, there are at most two, and thus exactly two, unbordered conjugates of f_n. □

Next we use the previous lemma to count the exact number of distinct frames in Fibonacci words. The number of distinct frames in the Fibonacci word f_n for small values of n is given in Table 2.

n	0	1	2	3	4	5	6	7	8	9	10
$F(f_n)$	0	0	0	0	1	3	5	6	8	10	12

Table 2: The number of distinct frames in Fibonacci words.

In the sequel, denote the first unbordered conjugate of f_n by $\hat{f}_n = af_{n-2}f_{n-1}^{\bullet}$, where $f_{n-1} = f_{n-1}^{\bullet}a$, and the second unbordered conjugate by $\tilde{f}_n = af_n^{\bullet}$, where $f_n = f_n^{\bullet}a$.

Theorem 1. *Let $n \geq 7$. The set of frames of f_n is*

$$\{f_1^2\} \cup \{\hat{f}_i^2 \mid i = 2, 3, \ldots, n-3\} \cup \{\tilde{f}_i^2 \mid i = 2, 3, \ldots, n-4\}.$$

In particular, we have $F(f_n) = 2(n-4)$.

Note that in the above n is circa $\log_{\phi}(F_n)$, where $\phi \approx 1.618$ denotes the golden number.

Proof. It is easy to verify that the statement holds for $n = 7$. Let $n > 7$ and assume that the claim holds for $n-1$. We prove that the Fibonacci word f_n

contains exactly two new frames, namely the frames $\hat{f}_{n-3}$ and $\tilde{f}_{n-4}$, which do not occur in f_{n-1}. This proves the claim.

(A) Let $\hat{f}$ denote the first unbordered conjugate of the Fibonacci word f_{n-3}. We show that the frame $\hat{f}\hat{f}$ is a factor of f_n, but it is not a factor of f_{n-1}. We have $f_{n-3} = f_{n-4}f_{n-5}$, where $f_{n-4} = f^{\bullet}_{n-4}a$ for a letter a, and $\hat{f} = af_{n-5}f^{\bullet}_{n-4}$. Hence,

$$\begin{aligned} f_n &= f_{n-1}f_{n-2} = f_{n-2}f_{n-3}f_{n-3}f_{n-4} \\ &= f_{n-2}(f_{n-4}f_{n-5})(f_{n-4}f_{n-5})f_{n-4} \\ &= f_{n-2}(f^{\bullet}_{n-4}af_{n-5})(f^{\bullet}_{n-4}af_{n-5})f^{\bullet}_{n-4}a \\ &= f_{n-2}f^{\bullet}_{n-4}(af_{n-5}f^{\bullet}_{n-4})(af_{n-5}f^{\bullet}_{n-4})a \\ &= f_{n-2}f^{\bullet}_{n-4}\hat{f}\hat{f}a, \end{aligned}$$

and therefore the frame $\hat{f}\hat{f}$ is a factor of f_n.

Assume now that $\hat{f}^2$ is a factor of f_{n-1}. Recall that $f_{n-4} = f^{\bullet}_{n-4}a$ and $f_{n-3} = f^{\bullet}_{n-4}af_{n-5}$. We have

$$\begin{aligned} f_{n-1} &= f_{n-2}f_{n-3} = f_{n-3}f_{n-4}f_{n-3} \\ &= f^{\bullet}_{n-4}af_{n-5}f^{\bullet}_{n-4}af_{n-3} \\ &= f^{\bullet}_{n-4}\hat{f}af_{n-3}. \end{aligned}$$

The word $\hat{f}$ is unbordered, and hence it does not overlap with itself. It follows that either (1) $\hat{f}^2 = af_{n-3}$ or (2) $\hat{f}$ is a prefix of af_{n-3}. In Case (1) we have $|f_{n-3}| = 1$. Then $\hat{f}^2 = aa$ is a suffix of f_{n-1}, which is impossible. In Case (2), we have $\hat{f} = af^{\bullet}_{n-3} = af^{\bullet}_{n-4}af^{\bullet}_{n-5}$, where $f_{n-3} = f^{\bullet}_{n-3}b$ and $f_{n-5} = f^{\bullet}_{n-5}b$ for a letter b. Then $af_{n-5}f^{\bullet}_{n-4} = \hat{f} = af^{\bullet}_{n-4}af^{\bullet}_{n-5}$, We thus have a border $af^{\bullet}_{n-5}$ in $\hat{f}$ contradicting the fact that $\hat{f}$ is unbordered.

(B) Let now $\tilde{f}$ denote the second unbordered conjugate of the Fibonacci word f_{n-4}. We show that the frame $\tilde{f}\tilde{f}$ is a factor of f_n, but it is not a factor of f_{n-1}. We have $f_{n-4} = f^{\bullet}_{n-4}a = f_{n-5}f^{\bullet}_{n-6}a$ and $f_{n-6} = f^{\bullet}_{n-6}a$ for a letter a.

Thus $\tilde{f} = af^{\bullet}_{n-4} = af_{n-5}f^{\bullet}_{n-6}$. Hence,

$$\begin{aligned}
f_n &= f_{n-1}f_{n-2} = f_{n-2}f_{n-3}f_{n-3}f_{n-4} \\
&= (f_{n-3}f_{n-4})(f_{n-4}f_{n-5})(f_{n-4}f_{n-5})f_{n-4} \\
&= (f_{n-3}f^{\bullet}_{n-4}a)(f^{\bullet}_{n-4}af_{n-5})(f_{n-5}f_{n-6}f_{n-5})f_{n-4} \\
&= (f_{n-3}f^{\bullet}_{n-4}a)(f^{\bullet}_{n-4}af_{n-5})(f_{n-6}f_{n-7}f_{n-6}f_{n-5})f_{n-4} \\
&= f_{n-3}f^{\bullet}_{n-4}(af^{\bullet}_{n-4})(af_{n-5}f^{\bullet}_{n-6}a)f_{n-7}f_{n-6}f_{n-5}f_{n-4} \\
&= f_{n-3}f^{\bullet}_{n-4}\tilde{f}\tilde{f}af_{n-7}f_{n-6}f_{n-5}f_{n-4},
\end{aligned}$$

and therefore, also in this case, the frame $\tilde{f}\tilde{f}$ is a factor of f_n.

Assume now that $\tilde{f}^2$ is a factor of f_{n-1}. We have

$$\begin{aligned}
f_{n-1} &= f_{n-2}f_{n-3} = f_{n-3}f_{n-4}f_{n-4}f_{n-5} \\
&= f_{n-3}f^{\bullet}_{n-4}af^{\bullet}_{n-4}af_{n-5} \\
&= f_{n-3}f^{\bullet}_{n-4}\tilde{f}af_{n-5}.
\end{aligned}$$

The word f_{n-3} does not end with a, since the last letters alternate in Fibonacci words (and f_{n-4} does end with a). The word $\tilde{f}$ is unbordered, and hence it does not overlap with itself. Therefore $\tilde{f}^2$ must either (1) occur in the prefix $f_{n-3}f^{\bullet}_{n-4}$ or (2) we have $\tilde{f} = af_{n-5}$. The latter case $af^{\bullet}_{n-4} = af_{n-5}$ never occurs if $n > 7$. Consider then case (1). The first occurrence of $\tilde{f}$ must be inside $f_{n-3} = f_{n-4}f_{n-5}$, but $|\tilde{f}| > |f_{n-5}|$ and $\tilde{f} = af^{\bullet}_{n-4}$ does not overlap with $f_{n-4} = f^{\bullet}_{n-4}a$ from the right (as it is unbordered) unless the overlap is the single letter a. In this case the suffix af_{n-5} of f_{n-3} is too short to contain $\tilde{f}$ if $n > 7$.

Now it remains to show that there are no other new frames in f_n. Assume that u^2 is a frame in f_n such that $|u| \geq 2$. By Lemma 2 and Lemma 3, we know that $u = \hat{f}_i$ or $u = \tilde{f}_i$ for some $i \geq 2$. According to case (B), the square of the second unbordered conjugate of f_{n-3} occurs for the first time in f_{n+1}. Similarly, by the cases above, the squares of the unbordered conjugates of f_{n-2} do not yet occur in f_n. Moreover, the squares of longer unbordered conjugates of Fibonacci words are too long to occur in f_n, and 11 is never a factor of a Fibonacci word. This proves the claim. □

4. Upper bound for distinct frames

Recall that $S(n)$ denotes the minimum number of occurrences of short frames (00, 11, 0101, 1010, 001001, 110110, 011011, 100100) that a word of

length n can contain. Table 1 suggests that $S(n) = \lceil n/2 \rceil - 2$. In this section we prove that the function $S(n)$ will not develop more exotic. For two words u and v, let $u \wedge v$ denote their longest common prefix.

In this section a word w is called *minimal* if $S(w) = S(|w|)$. First, let us show that in a minimal word w there are no short frames of length six.

Lemma 4. *Let w be a word containing the minimum number of occurrences of short frames. If u^2 is a short frame in w, then $u \in \{0, 1, 01, 10\}$.*

Proof. First, we show that a minimal word does not contain frames 100100 and 011011. Let $w = zuuy$ where the indicated occurrence of the square uu is the last one in w such that $u = 100$ or $u = 011$. Moreover, let w be a word such that the prefix z is as long as possible. In other words, if w' is another minimal word of length $|w|$, then the last occurrences of the frames 100100 and 011011 begin before position $|z| + 2$.

By symmetry, we can assume that $u = 100$. The word y is not empty, since otherwise $z100101$ would contain fewer short frames than w. Namely, the short frames 100100 and 00 end at the last position of $z100100$, but only one short frame 0101 ends at the last position of $z100101$. If $y = 1x$, then compare $w = z1001001x$ with $w' = z1001101x$. After the common prefix $w \wedge w' = z1001$, there are three frames (00, 100100 and 001001) ending in w but only one frame (11) ending in w' before x. By the minimality of w, the word w' must have at least two more short frames than w that end in the common suffix x. This implies that the prefix of x is 10. However, in this case the word w' has as many short frames as the minimal word w but 011011 occurs at the position $|z| + 3$. This is a contradiction. Hence, the suffix y begins with 0.

Now compare $w = z100100y$ with $\hat{w} = z100101y^c$, where y^c is the complement word of y, i.e., obtained by changing 0s and 1s. Since $y = 0x$ for some x, we have $w = z1001000x$ and $\hat{w} = z1001011x^c$. Now after the common prefix $w \wedge \hat{w} = z10010$ there are three frames, 00 twice and 100100 once, ending in w before x but only two frames, 0101 and 11, ending in $\hat{w}$ before x^c. By the minimality of w, there are more short frames ending in the suffix x^c of $\hat{w}$ than there are short frames ending in the suffix x of w. This implies that x begins with 100 and (x^c begins with 011). No matter how the suffix x continues, the word $\hat{w} = z1001011x^c$ has as many short frames as the minimal word $w = z1001000x$ but 011011 occurs at the position $|z| + 5$ of $\hat{w}$. Again, we have obtained a contradiction.

Hence, we have shown that minimal words do not contain frames 100100 and 011011. Moreover, this implies that there are no frames 001001 and 110110 in a minimal word w, since otherwise 100100 and 011011 would occur in the reversal w^R of w, which is also minimal. □

Lemma 5. *We have* $S(n) = \lceil n/2 \rceil - 2$ *for* $n \geq 3$.

Proof. We show first that $S(n) \geq \lceil n/2 \rceil - 2$. Note that the claim holds for $n \leq 8$. Let w be a word of length n for $n \geq 8$ such that $S(w) = S(n)$. We prove that for all u with $|u| = 2$, we have $S(wu) > S(w)$. The inequality follows from this.

Assume contrary to this that for some u with $|u| = 2$, we have $S(wu) = S(w)$. By symmetry, we can assume that w ends in the letter 0. Then $S(w0) > S(w)$, and therefore we need to consider only the cases where 1 is a prefix of the extending word u.

Case 1: Let $w = v10$ for some v. Now $wu = v1010$ or $wu = v1011$, and in both cases at least one new occurrence of a short frame is created, either 1010 or 11.

Case 2: Let $w = v00$ for some v. Since w is minimal, the word $w' = v01$ must contain at least $S(w)$ short frames, and since the last 00 is destroyed there exists a short frame f at the end of w'. We have two possibilities: $f = 0101$ and $f = 001001$. Since $S(w')$ is necessarily minimal, the case $f = 001001$ is impossible by Lemma 4. This means that w ends with 0100. Since u begins with 1 and $001001 \in S$, we must have $w = x10100$ (and $w' = x10101$) where $v = x101$. We compare w with the word $\hat{w} = x10110$. Now $w \wedge \hat{w} = x101$ of length $n - 2$, and there are equally many short frames ending in this portion of the two words. Since there are two short frames (00 and 1010) ending after $w \wedge \hat{w}$ in w, the minimality of $S(w)$ implies that in addition to the frame 11 there must be another short frame ending after $w \wedge \hat{w}$ in $\hat{w}$. The only possibility is that $\hat{w} = y110110$ where $x = y1$. Since $S(\hat{w}) = S(w)$ is again minimal, this contradicts with Lemma 4.

Finally we show that there is an equality in $S(n) = \lceil n/2 \rceil - 2$. The proof of this uses the same kind of case analysis as that for the inequality. Again, let w be a word of length n such that $S(w) = S(n)$. By Lemma 4, no two short frames are suffixes of w. Hence, it suffices to show that the last and the second last positions of w cannot both end in a short frame.

Assume to the contrary that if $w = va$ for $a \in \{0, 1\}$ then both w and v have a suffix from the set S. By Lemma 4, we need to consider two cases.

(a) Let $w = x000$. Thus both of the short frames are 00. Compare w with $w' = x001$ which must end in the small frame 001001 by the choice of w. Now w' contains as many frames as w. Thus, $S(w')$ is minimal. Since 001001 is a factor of w', this contradicts with Lemma 4.

(b) Let $w = x01010$. In this case w has a suffix 1010 and the second to last position ends necessarily in 0101. Compare w with $w' = x01101$. Here $w \wedge w' = x01$, after which w ends by two short frames, and hence also w' must end by at least two short frames. The frame 11 ends after the prefix $x01$ and another frame ending in w' after the common prefix is necessarily 110110. Then we have $w = y1101010$, where $x = y11$. Compare w with $\hat{w} = y1100110$. Now $w \wedge \hat{w} = y110$ after which w has an ending of three short frames (twice 1010 and once 0101) while $\hat{w}$ has an ending of only two short frames (00 and 11). This is again a contradiction.

The cases are exhausted, and hence $S(n) = \lceil n/2 \rceil - 2$. □

Note that, despite Lemma 4, the frames of length six are needed in the definition of the set S in order to obtain Lemma 5. Namely, the word $(100)^n$ of length $3n$ contains only n occurrences of short frames of length at most four (more precisely, n frames 00). However, we have $S((100)^n) = 3(n-1)$, which is far from the minimal value $S(3n) = \lceil 3n/2 \rceil - 2$ of occurrences of short frames.

Theorem 2. *A binary word w of length n can have at most $\lfloor n/2 \rfloor + 8$ different frames.*

Proof. Given $i = 2, 3, \ldots, n-1$, there can be at most one frame with the midpoint positioned at i. Indeed, if there are two squares uu and vv aligned after the first occurrences of u and v, then the shorter is a border of the larger. Hence there can be at most $n-2$ frames in w. By Lemma 5, there are at least $\lceil n/2 \rceil - 2$ occurrences of short frames, and a short frame can be counted only once. There are eight elements in S, and thus there are at most $n - 2 - (\lceil n/2 \rceil - 2 - 8) = \lfloor n/2 \rfloor + 8$ different frames. □

5. Lower bound for distinct frames

Consider the *Thue–Morse* words obtained by iterating the morphism $\mu\colon \{0,1\}^* \to \{0,1\}^*$ defined by $\mu(0) = 01$ and $\mu(1) = 10$. Let $\tau_i = \mu^i(0)$ for $i \geq 0$. Hence $|\tau_i| = 2^i$, and, e.g., $\tau_0 = 0$, $\tau_1 = 01$, $\tau_2 = 0110$, and $\tau_3 = 01101001$. Always, τ_i is a prefix of τ_{i+1}. Hence, there exists the infinite

Thue-Morse word $\tau = \lim_{i\to\infty} \tau_i = \lim_{i\to\infty} \mu^i(0)$. Now τ_i is a prefix of length 2^i of τ. Moreover, for $i > 0$, let $\hat{\tau}_i = \mu^{i-1}(011)$, i.e., the prefix of length $3 \cdot 2^{i-1}$ of τ.

The Thue–Morse words are overlap-free, see [22], and they have only a few frames. Indeed, the following result has been proven by Pansiot [23] and Brlek [24].

Lemma 6. *The squares of τ are all of the form $\mu^k(00)$, $\mu^k(11)$, or $\mu^k(010010)$, $\mu^k(101101)$ for some k.*

In particular, we have

Lemma 7. *The frames in τ are among the words* 00, 11, 0101 *and* 1010.

However, it was shown by Harju and Nowotka [25], that every other conjugate of τ_i and $\hat{\tau}_i$ is unbordered.

Lemma 8. *The word τ_i has 2^{i-1} unbordered conjugates and $\hat{\tau}_i$ has $3 \cdot 2^{i-2}$ unbordered conjugates.*

Let $\zeta\colon \{0,1\}^+ \to \{0,1\}^+$ be a mapping such that

$$\zeta(w) = waa,$$

where a is the last letter of w. We show that also $\zeta(\tau_i)$ and $\zeta(\hat{\tau}_i)$ have plenty of unbordered conjugates.

Lemma 9. *The word $\zeta(\tau_i)$ has 2^{i-1} unbordered conjugates and $\zeta(\hat{\tau}_i)$ has $3 \cdot 2^{i-2}$ unbordered conjugates.*

Proof. Let w be one of the words τ_i or $\hat{\tau}_i$ ending with $a \in \{0,1\}$. Consider a bordered conjugate of $\zeta(w)$ with minimal border u. In other words, the conjugate is of the form $u_1 x u_2$, where $u_1 = u_2 = u$ and x is some word. We show that this conjugate is either awa, aaw or a conjugate such that by deleting the two new letters a added by the mapping ζ we obtain a bordered conjugate of w. In the sequel these new letters are written in boldface.

The minimal border u of the conjugates $\mathbf{a}w\mathbf{a}$ and $\mathbf{aa}w$ is clearly a. Moreover, if the two new letters $\mathbf{a}$ added by the mapping ζ occur in x, then we may delete these letters and obtain one of the bordered conjugates of w. The same holds if x ends with $a\mathbf{a}$ and u_2 starts with $\mathbf{a}$. Also, observe that aaa is not a factor of the border u. Namely, no conjugate of $\zeta(w)$ can contain two

non-overlapping factors aaa, since w is overlap-free and it begins and ends with two distinct letters. Hence, let us assume that at least one $\mathbf{a}$ is a factor of the border u and $a\mathbf{aa}$ is not. It remains two cases to consider.

First, assume that u_1 ends with $a\mathbf{a}$ and, consequently, u_2 ends with aa and $u = u_1 = u_2$ begins with $b \neq a$. Note that $|u| = 2$ is impossible. This implies that $\zeta(w) = x'u_2u_1\mathbf{a} = x'(bu'aa)(bu'a\mathbf{a})\mathbf{a}$ for some words x' and u'. Since w consists of blocks 01 and 10, we conclude that $x'(bu'aa)$ must have odd length. Since $w = x'(bu'aa)bu'a$ has even length, the length of $u = u_1 = u_2$ must be even. Hence, the words u_1 and u_2 start inside a block ab and $w = x''a(bu'aa)bu'a = x''vv$, where $v = abu'a$. This is impossible, since no square is a suffix of w by Lemma 6.

Next, consider the case where u_2 begins with $\mathbf{aa}$. Then u_1 begins with aa and $u = u_1 = u_2$ ends with b. Note that $|u| = 2$ is here impossible. Since u_1 begins with aa, the last letter of u_2 and the first letter of u_1 must form a block ba in the word $\zeta(w)$. Hence, the length of u must be odd. Therefore, by considering the block structure of w, we conclude that u_1 must end with the block ab, which implies that u_2 ends with bab and, consequently, u_1 ends with $abab$. Hence, we have $\zeta(w) = u'ababu_1x\mathbf{aa}$ for some word u'. Since u_1 begins with a, the word w has a factor $ababa$ contradicting the overlap-freeness of w.

Hence, we have proved that the number of unbordered conjugates of $\zeta(w)$ is exactly the number of unbordered conjugates of w. By Lemma 8, this means that in $\zeta(\tau_i)$ of length $2^i + 2$ there are 2^{i-1} unbordered conjugates, and in $\zeta(\hat{\tau}_i)$ of length $3 \cdot 2^{i-1} + 2$ there are $3 \cdot 2^{i-2}$ unbordered conjugates. □

Using Lemma 9 we obtain a lower bound for $F(n)$.

Theorem 3. *We have $F(n) > \frac{7}{30}n - \epsilon$ for any positive ϵ and infinitely many n.*

Proof. Consider first the word

$$u_k = \hat{\tau}_1^2\tau_2^2\hat{\tau}_2^2 \cdots \tau_{k-1}^2\hat{\tau}_{k-1}^2\tau_k^2\tau_k$$

of length

$$\begin{aligned} |u_k| &= 2\sum_{i=1}^{k-1} 3 \cdot 2^{i-1} + 2\sum_{i=2}^{k} 2^i + 2^k \\ &= 2^{k+3} - 14. \end{aligned}$$

Now $\hat{\tau}_1 = 011$ has two unbordered conjugates, but otherwise $\hat{\tau}_i$ has $3 \cdot 2^{i-2}$ unbordered conjugates by Lemma 8. Similarly, the Thue-Morse word τ_i has 2^{i-1} unbordered conjugates. Also, when $k \geq 4$, u_k has the four short frames 00, 11 and 0101, 1010. Therefore,

$$\begin{aligned} F(u_k) &\geq 4 + 2 + \sum_{i=2}^{k-1} 3 \cdot 2^{i-2} + \sum_{i=2}^{k} 2^{i-1} \\ &= 7 \cdot 2^{k-2} + 1. \end{aligned}$$

Consider next the word

$$v_k = \zeta(\hat{\tau}_1)^2 \zeta(\tau_2)^2 \zeta(\hat{\tau}_2)^2 \cdots \zeta(\tau_{k-1})^2 \zeta(\hat{\tau}_{k-1})^2 \zeta(\tau_k)^2 \tau_k$$

of length $|u_k| + 8(k-1) = 2^{k+3} + 8k - 22$. By Lemma 9, we have at least the same number of frames in v_k as in u_k, i.e., $F(v_k) \geq 7 \cdot 2^{k-2} + 1$.

Now let us combine these two words. Notice that the reversal $(v_k)^R$ of the word v_k has the same length and contains the same number of frames as v_k. Moreover, the Thue-Morse word τ_i is a palindrome for even values of i. Hence, we have

$$\begin{aligned} (v_{2k})^R &= \tau_{2k}^R (\zeta(\hat{\tau}_1)^2 \zeta(\tau_2)^2 \zeta(\hat{\tau}_2)^2 \cdots \zeta(\tau_{2k-1})^2 \zeta(\hat{\tau}_{2k-1})^2 \zeta(\tau_{2k})^2)^R \\ &= \tau_{2k} (\zeta(\hat{\tau}_1)^2 \zeta(\tau_2)^2 \zeta(\hat{\tau}_2)^2 \cdots \zeta(\tau_{2k-1})^2 \zeta(\hat{\tau}_{2k-1})^2 \zeta(\tau_{2k})^2)^R \end{aligned}$$

Consider now the word

$$\begin{aligned} w_k &= u_{2k} (\zeta(\hat{\tau}_1)^2 \zeta(\tau_2)^2 \zeta(\hat{\tau}_2)^2 \cdots \zeta(\tau_{2k-1})^2 \zeta(\hat{\tau}_{2k-1})^2 \zeta(\tau_{2k})^2)^R \\ &= \hat{\tau}_1^2 \tau_2^2 \cdots \hat{\tau}_{2k-1}^2 \tau_{2k}^2 \tau_{2k} (\zeta(\hat{\tau}_1)^2 \zeta(\tau_2)^2 \cdots \zeta(\hat{\tau}_{2k-1})^2 \zeta(\tau_{2k})^2)^R. \end{aligned}$$

This word contains all frames of u_{2k} and $(v_{2k})^R$. Note that 00, 11, 0101 and 1010 are common to both words. However, frame uu, where u is a conjugate of τ_i or $\hat{\tau}_i$ differs from vv, where v is a conjugate of $\zeta(\tau_i)$ or $\zeta(\hat{\tau}_i)$, since $|v|_0 \neq |v|_1$ but $|u|_0 = |u|_1$. Hence, we have

$$F(w_k) \geq 14 \cdot 2^{2k-2} - 2.$$

The length of w_{2k} is

$$\begin{aligned} |w_k| &= |u_{2k}| + |v_{2k}| - 2^{2k} \\ &= (2^{2k+3} - 14) + (2^{2k+3} + 8 \cdot 2k - 22) - 2^{2k} \\ &= 15 \cdot 2^{2k} + 16k - 36. \end{aligned}$$

Therefore

$$\frac{F(w_k)}{|w_k|} \geq \frac{14 \cdot 2^{2k-2} - 2}{15 \cdot 2^{2k} + 16k - 36}.$$

Since

$$\lim_{k\to\infty} \frac{14 \cdot 2^{2k-2} - 2}{15 \cdot 2^{2k} + 16k - 36} = \frac{14}{15 \cdot 4} = \frac{7}{30},$$

this proves the claim. □

6. Minimum number of occurrences of frames

Recall that $M(w)$ is the number of occurrences of frames in a binary word w and $M(n) = \min\{M(w) \mid |w| = n\}$. In this section we are interested in finding words of given length having as few frames as possible. Note that maximizing the value $M(w)$ for words of given length n is easy. For example, the word $w = 0^n$ contains $n-1$ frames 00 and gives the maximum $M(w) = n - 1$.

As in the previous section, consider the Thue–Morse word τ and its prefixes τ_i. Now we may count the number of occurrences of frames in a prefix of even length of the Thue-Morse word.

Lemma 10. *Let w be a prefix of the Thue-Morse word τ of even length $n > 2$. Then*

$$M(w) = \frac{|w|}{2} - 1.$$

Proof. For $k \geq 2$, we have $\tau_{k+1} = \mu^k(0)\mu^k(1)$, where $M(\mu^k(0)) = M(\mu^k(1))$, and the only new frame created between the parts $\mu^k(0)$ and $\mu^k(1)$ is either 11 or 1010. Indeed, $\mu^k(1)$ begins with 1001 and if k is odd, $\mu^k(0)$ ends in 1001 and if k is even, $\mu^k(0)$ ends in 0110. Hence $M(\tau_{k+1}) = 2M(\tau_k) + 1$. Since $M(\tau_2) = 1$, this gives us

$$M(\tau_k) = 2^{k-1} - 1 = |\tau_k|/2 - 1 \tag{1}$$

for all $k \geq 2$. The prefixes w of even length of τ_k are of the form

$$w \in \{u0110, u0101, u1001, u1010\},$$

and in these cases $M(u01) = M(w)-1$ and $M(u10) = M(w)-1$ by Lemma 7. Hence, the claim follows from (1). □

Next we modify the prefixes of the Thue-Morse word to obtain a word of length n that has the minimum number of occurrences of frames. We show that $M(n) = S(n)$, where $S(n)$ is the minimum number of occurrences of short frames defined in Section 2. Note that our words w satisfying $M(w) = M(n)$ contain only short frames 00, 11, 0101 and 1010. Occasionally, we write a dot $(u.v)$ to emphasize a decomposition of the word.

Theorem 4. *For $n \geq 3$, we have*

$$M(n) = \left\lceil \frac{n}{2} \right\rceil - 2. \tag{2}$$

Proof. It is clear that $M(n) \geq S(n) = \left\lceil \frac{n}{2} \right\rceil - 2$, since each word of length n has at least $S(n)$ short frames. As in Section 3 we denote $w^\bullet = v$ for $w = va$ where $a \in \{0, 1\}$. Moreover, if v is a prefix of τ of length m, then we denote

$$\tau = v\tau^{(m)}.$$

We say that a word w is *fit* if it satisfies (2) for $n = |w|$. The words 0101 and 1010 are called *flips* in the Thue-Morse word. A *flip word* is a word that ends with a flip. We note that if w is a prefix of τ such that $|w| \equiv 6 \pmod 8$ then w is a flip word. (There are also other flips in τ, but we do not use them.) We have by Lemma 7 that if w is a flip word then

$$M(w^\bullet) = M(w) - 1. \tag{3}$$

We reduce the claim to the even cases of Lemma 10 in four steps by which we obtain fit words.

(A) Let $\alpha = 1\tau$, i.e.,

$$\alpha = 1.0110.1001.1001\ldots.$$

If w is a prefix of τ of even length, then $1w$ is fit. For this one needs only to show that there are no frames as a prefix of α.

Assuming the opposite, we conclude that $1w$ begins with the frame $(101101z00)^2$, where z is some word. This implies that w, which is a prefix of τ, begins with the the square $(01101z001)^2$. However, by Lemma 6, no square is a prefix of the Thue-Morse word.

Hence, if $|w| = n$ is even, then by Lemma 10, we have

$$M(1w) = \frac{n}{2} - 1 = \frac{n+2}{2} - 2 = \left\lceil \frac{|1w|}{2} \right\rceil - 2.$$

Thus, for all odd lengths, there exists a fit word. Also, by the above, if $|1w| \equiv 7 \pmod 8$ then $1w$ is a flip word. Hence, by (3), we have fit words $1w^\bullet$ of all lengths $n \equiv 6 \pmod 8$.

(B) Let $\beta = \tau^{(3)}$, i.e.,

$$\beta = 0.1001.1001\ldots.$$

Now, $\tau = 011\tau^{(3)}$. If w is a prefix of β of odd length, then it is fit. Indeed, $011w$ is an even length prefix of τ, which contains $(|w|+3)/2-1$ frames by Lemma 10. Since the first two frames 11 and 1010 of τ do not occur in w, we have by Lemma 7 that $M(w) = M(011w) - 2 = \lceil |w|/2 \rceil - 2$.

Therefore, by the above, if $|w| \equiv 3 \pmod 8$ then w is a flip word. Hence, we have fit words $w^\bullet$ of all lengths $n \equiv 2 \pmod 8$.

(C) Let $\gamma = 1\tau^{(6)}$, i.e.,

$$\gamma = 101.1001.0110\ldots.$$

Now, $\tau = 011010\tau^{(6)}$. If w is a prefix of $\tau^{(6)}$ of even length, then $1w$ is fit. Namely, there are no frames in the beginning of $1w$. Otherwise, the frame would be of the form $(1011z00)^2$. By the structure of the Thue-Morse word, this implies that the square $(011z001)^2$ is a prefix of w and therefore a square in the Thue-Morse word τ. Since it begins with zero, it must be of the form $\mu^k(00)$ or $\mu^k(010010)$ by Lemma 6. We notice that $\mu^k(00)$ and $\mu^k(010010)$ are not prefixes of $\tau^{(6)}$ for $k = 0, 1, 2$. Moreover, $\tau^{(6)} = 011001\cdots$, but $\mu^k(00)$ and $\mu^k(010010)$ begin with 011010 for $k \geq 3$. Hence, adding 1 in front of w does not increase the number of frames. By Lemma 7, we know that the three first frames of $\tau = 011010\tau^{(6)}$ are 11, 1010 and 00. Subtracting these occurrences, we conclude by Lemma 10 that

$$M(1w) = M(011010w) - 3 = \frac{|w|+6}{2} - 4 = \left\lceil \frac{|1w|}{2} \right\rceil - 2.$$

Therefore, by the above, if $|1w| \equiv 1 \pmod 8$ then $1w$ is a flip word. Hence we have fit words $1w^\bullet$ of all lengths $n \equiv 0 \pmod 8$.

(D) Let $\delta = 010\tau^{(12)}$, i.e.,

$$\delta = 010.0110.1001\ldots.$$

Now, $\tau = 011010011001\tau^{(12)}$. If w is a prefix of $\tau^{(12)}$ of even length, then $010w$ is fit. In order to prove this, we show that $M(010w) = M(w)$.

First, we notice that 00 is the only frame in the beginning of $0w$. As in the previous cases, if there is a longer frame in the beginning of $0w$, then it is of the form $(001z1)^2$. Since $\tau^{(12)}$ consists of blocks 01 and 10, we conclude that $001z1$ must have even length and, consequently, $\tau^{(12)}$ begins with the square $(01z10)^2$. By Lemma 6, this square is either $\mu^k(00)$ or $\mu^k(010010)$ for some k. For $0 \leq k \leq 3$, we can easily check that the squares $\mu^k(00)$ or $\mu^k(010010)$ are not prefixes of $\tau^{(12)}$. Moreover, for $k \geq 4$, the words $\mu^k(00)$ or $\mu^k(010010)$ begin with 0110.1001.1, whereas $\tau^{(12)}$ begins with 0110.1001.0. Hence, we have proved that $M(0w) = M(w) + 1$.

Assume next that $10w$ begins with a frame. This frame must be of the form $(1001z00)^2$. Hence, 00100 becomes a factor of the Thue-Morse word, which is clearly impossible. Thus, we have $M(10w) = M(0w) = M(w) + 1$.

Similarly, if $010w$ begins with a frame, this frame is of the form $(010.0110.10z)^2$. Now 010011010 must occur in $\tau \in \{0110, 1001\}^*$, which is a contradiction. This gives us $M(010w) = M(10w) = M(w) + 1$.

By Lemma 7, we find out that $M(w) = M(0110.1001.1001w) - 6$. By Lemma 10, this gives us

$$M(010w) = M(w) + 1 = \frac{|w| + 12}{2} - 6 = \frac{|w| + 4}{2} - 2 = \left\lceil \frac{|010w|}{2} \right\rceil - 2.$$

If $|w| \equiv 2 \pmod 8$, then $0110.1001.1001w$ is a prefix of τ of length $8k+6$. Thus, $|010w| \equiv 5 \pmod 8$ and $010w$ is a flip word. Hence we have fit words $010w^\bullet$ of all lengths $n \equiv 4 \pmod 8$. This proves the claim.

□

7. Conclusions

We have shown that the minimum number of occurrences of frames in a binary word of length n is $M(n) = \lceil \frac{n}{2} \rceil - 2$ and the maximum number $F(n)$ of different frames is at most $\lfloor n/2 \rfloor + 8$. On the other hand, $F(n) > \frac{7}{30}n - \epsilon$ for any positive ϵ and infinitely many n. Moreover, for words of length n, the maximum number of occurrences of frames is trivially $n-1$ and the minimum number of distinct frames is three for $n > 18$. It will be a challenging task to minimize the gap between the upper bound and the lower bound of $F(n)$. We conjecture that $F(n) = n/4$ for n large enough.

[1] A. Thue, Über unendliche Zeichenreihen, Norske Vid. Skrifter I Mat.-Nat. Kl., Christiania 7 (1906) 1–22.

[2] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, Norske Vid. Skrifter I Mat.-Nat. Kl., Christiania 1 (1912) 1–67.

[3] A. S. Fraenkel, J. Simpson, How many squares can a string contain?, J. Combin. Theory Ser. A 82 (1) (1998) 112–120.

[4] L. Ilie, A simple proof that a word of length n has at most $2n$ distinct squares, J. Combin. Theory Ser. A 112 (1) (2005) 163–164.

[5] L. Ilie, A note on the number of squares in a word, Theoret. Comput. Sci. 380 (3) (2007) 373–376.

[6] A. S. Fraenkel, R. J. Simpson, How many squares must a binary sequence contain?, Electron. J. Combin. 2 (1995) Research Paper 2, approx. 9 pp. (electronic).

[7] T. Harju, D. Nowotka, Binary words with few squares, Bull. Eur. Assoc. Theor. Comput. Sci. EATCS (89) (2006) 164–166.

[8] N. Rampersad, J. Shallit, M. Wang, Avoiding large squares in infinite binary words, Theoret. Comput. Sci. 339 (1) (2005) 19–34.

[9] G. Kucherov, P. Ochem, M. Rao, How many square occurrences must a binary sequence contain?, Electron. J. Combin. 10 (2003) Research Paper 12, 11 pp. (electronic).

[10] M. Crochemore, An optimal algorithm for computing the repetitions in a word, Inform. Process. Lett. 12 (1981) 244–250.

[11] J. Berstel, D. Perrin, Theory of Codes, Vol. 117 of Pure and Applied Mathematics, Academic Press Inc., Orlando, FL, 1985.

[12] W.-F. Chuan, Unbordered factors of the characteristic sequences of irrational numbers, Theoret.Comput.Sci. 205 (1998) 337–344.

[13] J. C. Costa, Biinfinite words with maximal recurrent unbordered factors, Theoret.Comput.Sci. 290 (2003) 2053–2061.

[14] J.-P. Duval, Relationship between the period of a finite word and the length of its unbordered segments, Discrete Math. 40 (1982) 31–44.

[15] A. Ehrenfeucht, D. M. Silberger, Periodicity and unbordered segments of words, Discrete Math. 26 (1979) 101–109.

[16] H. Morita, A. J. van Wijngaarden, A. J. Han Vinck, On the construction of maximal prefix-synchronized codes, IEEE Trans.Inform. Theory 42 (1996) 2158–2166.

[17] A. S. Fraenkel, J. Simpson, The exact number of squares in Fibonacci words, Theoret. Comput. Sci. 218 (1) (1999) 95–106, WORDS (Rouen, 1997).

[18] M. Lothaire, Algebraic Combinatorics on Words, Vol. 90 of Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 2002.

[19] D. Breslauer, T. Jiang, Z. Jiang, Rotations of periodic strings and short superstrings, J. Algorithms 24 (2) (1997) 340–353.

[20] J. Currie, K. Saari, Least periods of factors of infinite words., Theor. Inf. Appl. 43 (1) (2009) 165–178.

[21] T. Harju, D. Nowotka, Density of critical factorizations, Theor. Inf. Appl. 36 (2002) 315–327.

[22] M. Lothaire, Combinatorics on Words, Vol. 17 of Encyclopedia of Mathematics and its Applications, Addison-Wesley, 1983.

[23] J. J. Pansiot, The Morse sequence and iterated morphisms, Inform. Process. Lett. 12 (1981) 68–70.

[24] S. Brlek, Enumeration of factors in the Thue-Morse word, Disc. Appl. Math. 24 (1989) 83–86.

[25] T. Harju, D. Nowotka, Border correlation of binary words, J. Combin. Theory Ser. A 108 (2) (2004) 331–341.