Maximum Number of Distinct and Nonequivalent Nonstandard Squares in a Word[⋆]

Tomasz Kociumaka[1,⋆⋆], Jakub Radoszewski[1,⋆⋆⋆],
Wojciech Rytter[1,2], and Tomasz Waleń[1,†]

[1] Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw, Warsaw, Poland
[kociumaka,jrad,rytter,walen]@mimuw.edu.pl
[2] Faculty of Mathematics and Computer Science,
Copernicus University, Toruń, Poland

Abstract. The combinatorics of squares in a word depends on how the equivalence of halves of the square is defined. We consider Abelian squares, parameterized squares and order-preserving squares. The word uv is an Abelian (parameterized, order-preserving) square if u and v are equivalent in the Abelian (parameterized, order-preserving) sense. The maximum number of ordinary squares is known to be asymptotically linear, but the exact bound is still investigated. We present several results on the maximum number of distinct squares for nonstandard subword equivalence relations. Let $SQ_{\mathrm{Abel}}(n,k)$ and $SQ'_{\mathrm{Abel}}(n,k)$ denote the maximum number of Abelian squares in a word of length n over an alphabet of size k, which are distinct as words and which are nonequivalent in the Abelian sense, respectively. We prove that $SQ_{\mathrm{Abel}}(n,2) = \Theta(n^2)$ and $SQ'_{\mathrm{Abel}}(n,2) = \Omega(n^{1.5}/\log n)$. We also give linear bounds for parameterized and order-preserving squares for small alphabets: $SQ_{\mathrm{param}}(n,2) = \Theta(n)$ and $SQ_{\mathrm{op}}(n,O(1)) = \Theta(n)$. As a side result we construct infinite words over the smallest alphabet which avoid nontrivial order-preserving squares and nontrivial parameterized cubes (nontrivial parameterized squares cannot be avoided in an infinite word).

1 Introduction

Repetitions in words are a fundamental topic in combinatorics on words [2]. They are widely used in many fields, such as pattern matching, automata theory, formal language theory, data compression, molecular biology, etc. Squares, that is, words of the form uu, are one of the most commonly studied types of repetitions. An example of an infinite square-free word over a ternary alphabet, given by Thue [24], is considered to be the foundation of combinatorics on words.

[⋆] The authors thank anonymous referees for many helpful comments.

[⋆⋆] The author is supported by Polish budget funds for science in 2013-2017 as a research project under the 'Diamond Grant' program.

[⋆⋆⋆] The author receives financial support of Foundation for Polish Science.

[†] The author is supported by Iuventus Plus grant (IP2011 058671) of the Polish Ministry of Science and Higher Education.

If we allow other equivalence relations on words, several generalizations of the notion of square can be obtained. One such generalization are Abelian squares, that is, words of the form uv where the multisets of symbols of u and v are the same. Abelian squares were first studied by Erdős [10], who posed a question on the smallest alphabet size for which there exists an infinite Abelian-square-free word. The first example of such a word over a finite alphabet was given by Evdokimov [11], later the alphabet size was improved to five by Pleasants [23] and finally an optimal example over four-letter alphabet was shown by Keränen [20].

In this paper we consider Abelian squares and introduce squares based on two other known equivalence relations on words. The first is parameterized equivalence [1], in which two words u, v of length n over alphabets $\mathrm{Alph}(u)$ and $\mathrm{Alph}(v)$ are considered equal if one can find a bijection $f : \mathrm{Alph}(u) \to \mathrm{Alph}(v)$ such that $v[i] = f(u[i])$ for all $i = 1, \ldots, n$. The second model, order-preserving equivalence [6], assumes that the alphabets are ordered. Two words u, v of the same length are considered equivalent in this model if they are equal in the parameterized sense with f being an strictly increasing bijection. We define a parameterized square and an order-preserving square as a concatenation of two words that are equivalent in the parameterized and in the order-preserving sense, respectively. Another recently studied model, lying in between standard equality and Abelian equivalence, is k-Abelian equivalence [17]. However, we do not consider this model here. The nonstandard types of squares can be viewed as a part of nonstandard stringology; see [21,22].

Example 1. Consider the alphabet $\Sigma = \{1, 2, 3, 4\}$ with the natural order. Then 1213 1213 is a square, 1213 3112 is an Abelian square, 1213 4142 is a parameterized square, and 1213 1314 is an order-preserving square over Σ.

An important combinatorial fact about ordinary squares is that the maximum number of distinct squares in a word of length n is linear in terms of n. Actually this number is smaller than $2n - \Theta(\log n)$ [14,18,19]. This bound has found applications in several text algorithms [5] including two different linear-time algorithms computing all distinct squares [15,8]. A recent result shows that the maximum number of distinct squares in a labeled tree is asymptotically $\Theta(n^{4/3})$ [7]. Also some facts about counting distinct squares in partial words are known [3,4]. In this paper we attempt the same type of combinatorial analysis for nonstandard squares. In turns out that the results that we obtain depend heavily on which squares we consider distinct.

Let $SQ_{\mathrm{Abel}}(n, k)$, $SQ_{\mathrm{param}}(n, k)$ and $SQ_{\mathrm{op}}(n, k)$ denote respectively the maximum number of Abelian, parameterized and order-preserving squares in a word of length n over an alphabet of size k which are *distinct* as words. Moreover let $SQ'_{\mathrm{Abel}}(n, k)$, $SQ'_{\mathrm{param}}(n, k)$ and $SQ'_{\mathrm{op}}(n, k)$ denote the maximum number of Abelian, parameterized and order-preserving squares in a word of length n over an alphabet of size k which are *nonequivalent* in the Abelian, parameterized and order-preserving sense, respectively. We also use analogous notation, e.g., $SQ_{\mathrm{Abel}}(w)$, $SQ'_{\mathrm{Abel}}(w)$, for any word w. Our main results are the following:

- $SQ_{\mathrm{Abel}}(n, 2) = \Theta(n^2)$, $SQ'_{\mathrm{Abel}}(n, 2) = \Omega(n^{1.5}/\log n)$;

- $SQ_{\mathrm{op}}(n,k) = \Theta(n)$ and therefore $SQ'_{\mathrm{op}}(n,k) = \Theta(n)$ for $k = O(1)$;
- $SQ_{\mathrm{param}}(n,2) = \Theta(n)$ and therefore $SQ'_{\mathrm{param}}(n,2) = \Theta(n)$.

Example 2. Consider a Fibonacci word $Fib_5 = 0100101001001$.[3] It contains 5 Abelian squares of length 6: 010 010, 001 010, 010 100, 100 100, and 001 001, which are all distinct as words but are Abelian-equivalent. In total Fib_5 contains 13 distinct subwords which are Abelian squares. Hence, $SQ_{\mathrm{Abel}}(Fib_5) = 13$. On the other hand, Fib_5 contains only 5 Abelian-nonequivalent squares, with sample representatives: 0 0, 01 01, 001 010, 10010 10010, and 010010 100100. Hence, $SQ'_{\mathrm{Abel}}(Fib_5) = 5$. The value SQ' is usually much smaller than SQ, e.g., for Fib_{14} of length 987, $SQ'_{\mathrm{Abel}}(Fib_{14}) = 490$ and $SQ_{\mathrm{Abel}}(Fib_{14}) = 57796$. In general one can show that $SQ'_{\mathrm{Abel}}(Fib_k) = O(|Fib_k|)$. Abelian repetitions in Fibonacci words and Sturmian words were already studied in [13].

The second part of our paper can be viewed as an extension of the works of Thue [24], Evdokimov [11], Pleasants [23] and Keränen [20] on infinite square-free and Abelian-square-free words into the parameterized and order-preserving equivalence. As no square-free word of length larger than 1 exists, we consider words avoiding *nontrivial* nonstandard squares of length larger than 2. We present an infinite word over the minimum-size (ternary) alphabet avoiding nontrivial order-preserving squares. We also prove that there is no infinite word avoiding nontrivial parameterized squares, but there is one avoiding nontrivial parameterized cubes, that is, parameterized cubes of length greater than 3.

2 Bounds for Abelian Squares

For a word $w = w[1]\cdots w[n]$ we denote $|w| = n$. A *subword* of w is a word of the form $w[i]\cdots w[j]$ for $1 \le i \le j \le |w|$, which we denote by $w[i..j]$. A word is said to be *uniform* if all its letters are equal. A *block* (also known as a *run*) in a word is a maximal uniform subword.

In this section we restrict ourselves to the binary alphabet. First, we show a simple example which yields $SQ_{\mathrm{Abel}}(n,2) = \Theta(n^2)$. Afterwards we attempt an analysis of $SQ'_{\mathrm{Abel}}(n,2)$. Our main result is a lower bound of $\Omega(n^{1.5}/\log n)$. We also obtain an upper bound $O(nm)$ if the number of blocks is bounded by m.

A different proof of the following theorem was given independently by Fici [12].

Theorem 1. *$SQ_{\mathrm{Abel}}(n,2) = \Theta(n^2)$.*

Proof. Consider the word $u_k = 0^k10^k10^{2k}$ of length $4k+2$. It contains $\Theta(k^2)$ Abelian squares of the form $0^a10^b\ 0^{k-b}10^{a+2b-k}$ for all $0 \le a, b \le k$ and $a+2b \ge k$. Thus we obtain $SQ_{\mathrm{Abel}}(n,2) = \Theta(n^2)$ for $n = 4k+2$. If $n \bmod 4 \ne 2$, we pick the longest word u_k such that $|u_k| \le n$ and extend it with $n-|u_k| \le 3$ zeros. □

[3] Fibonacci words are defined as: $Fib_0 = 0$, $Fib_1 = 01$, $Fib_k = Fib_{k-1}Fib_{k-2}$ for $k \ge 2$.

2.1 Lower bound for $SQ'_{\mathbf{Abel}}(n, 2)$

For a word w and a letter c we denote the number of occurrences of c in w by $|w|_c$. The Parikh vector of a binary word w is $\mathcal{P}(w) = (|w|_0,\ |w|_1)$.

We say that (p, q) is *a square vector* in w if there exists an Abelian square u_1u_2 in w such that $\mathcal{P}(u_1) = \mathcal{P}(u_2) = (p, q)$. Then u_1u_2 is called a (p, q)-square. Let $SQV(w)$ denote the set of square vectors of w. Now $SQ'_{\text{Abel}}(n, 2)$ is the maximum number of different square vectors in a binary word of length n.

In the proof of the lower bound we require some number-theoretic tools. Erdős [9] investigated the problem of estimating the numbers:

$$P_k = |\{i \cdot j : 1 \leq i, j \leq k\}|.$$

It is known that $P_k = \Omega(k^2/\log k)$. Our auxiliary problem is similar, but instead of the ordinary multiplication $i \cdot j$ we consider an operation

$$i \otimes j \ = \ \textstyle\sum_{t=i}^{j}\ t = (i + j)(j - i + 1)/2.$$

We define

$$\text{Sums}(a, b) = |\{i \otimes j \ : a \leq i \leq j \leq b\}|.$$

Example 3. $\text{Sums}(2, 5) \ = \ \{2, 3, 4, 5, 7, 9, 12, 14\}$.

Lemma 1. $\text{Sums}(\lceil \frac{3}{4}k\rceil, k) \ = \ \Omega(k^2/\log k)$.

Proof. We use the following textbook fact:

Fact 1 ([16]). *Let* $\pi(x)$ *be the number of prime numbers in the range* $[1..x]$. *For any* $\varepsilon > 0$ *we have* $\pi((1 + \varepsilon)x) - \pi(x) = \frac{\varepsilon x}{\log x} + o(\frac{x}{\log x})$.

Let I_k denote the set of primes in the interval $\left[\lceil \frac{10}{12}k\rceil, \lfloor \frac{11}{12}k\rfloor\right]$ (this interval is a *middle third* of $\left[\lceil \frac{3}{4}k\rceil, k\right]$). Let

$$F_k = \{(i, j) \ : \ 0 \leq j - i < \lfloor \tfrac{k}{12}\rfloor - 1 \text{ and } \tfrac{i+j}{2} \in I_k\}.$$

Fact 1 implies that $|I_k| = \Omega(k/\log k)$, and consequently $|F_k| = \Omega(k^2/\log k)$. Note that $\{i \otimes j : (i, j) \in F_k\} \subseteq \text{Sums}(\lceil \frac{3}{4}k\rceil, k)$. Therefore it suffices to prove the following:

Claim. If $(i, j), (i', j') \in F_k$, $(i, j) \neq (i', j')$, then $i \otimes j \neq i' \otimes j'$.

However $i \otimes j = p \cdot (j - i + 1)$, $i' \otimes j' = p' \cdot (j' - i' + 1)$, for $p, p' \in I_k$. The claim follows from the primality of p, p' and the inequalities $j - i + 1, j' - i' + 1 \leq \min(p, p')$. □

In our construction a crucial role is played by *balanced* Abelian squares and *balanced* square vectors. A square vector (p, q) is called balanced if $p = q$, and a word w is called balanced if its Parikh vector is balanced. We define

$$\begin{aligned}
\text{neigh}^+((p, q), r) &= \{(p, q + t) \ : \ 0 \leq t \leq r\}, \\
\text{neigh}^-((p, q), r) &= \{(p, q - t) \ : \ 0 \leq t \leq r\}, \\
\text{neigh}((p, q), r) &= \text{neigh}^+((p, q), r) \cup \text{neigh}^-((p, q), r).
\end{aligned}$$

For $i \leq j$ let us define the following word of length $2 \cdot (i \otimes j)$:

$$\mathbf{w}_{i,j} = 0^i1^i0^{i+1}1^{i+1} \cdots 0^j1^j.$$

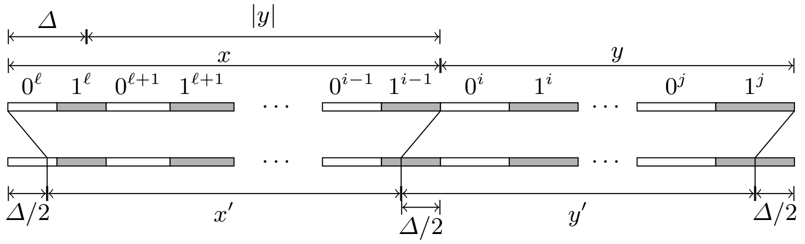

Fig. 1. Illustration of the proof of Lemma 2 — construction of balanced Abelian square $x'y'$.

Observation 1. *Let $w = \mathbf{w}_{i,j}$ and $\varDelta \in \{0, \ldots, i\}$. Then the subword $w[1 + \varDelta..|w| - \varDelta]$ is balanced.*

Let us take $\mathbf{w}_k = \mathbf{w}_{1,k}$. For example $\mathbf{w}_4 = 01001100011100001111$. Also, let $\mathbf{S}_k$ be a family of balanced vectors $\big\{(p,p) : p \in \mathrm{Sums}\,\big(\lceil \frac{3}{4}k \rceil, k\big)\big\}$.

Lemma 2. *If $k > 16$, then $\mathbf{S}_k \subseteq SQV(\mathbf{w}_k)$.*

Proof. Let $p = i \otimes j$ for i, j such that $\lceil \frac{3}{4}k \rceil \le i, j \le k$ and let $\ell < i$ be the largest index such that $\ell \otimes (i-1) \ge p$. Such an integer ℓ exists since for $k > 16$ we have $1 \otimes \big(\lceil \frac{3}{4}k \rceil - 1\big) \ge \lceil \frac{3}{4}k \rceil \otimes k$.

Consider the subwords $x = \mathbf{w}_{\ell,i-1}$, $y = \mathbf{w}_{i,j}$ of $\mathbf{w}_k$. If $|x| = |y|$, then we have just located a square xy with a square vector (p,p) and we are done. Otherwise, let $\varDelta = |x| - |y| > 0$; see Fig. 1. Note that $0 < \varDelta/2 < \ell$ and $|x|_0 = |x|_1 = p + \varDelta/2$. We modify x into x' by cutting away the first $\varDelta/2$ zeros and the last $\varDelta/2$ ones: $x' = x[\varDelta/2 + 1..|x| - \varDelta/2]$. Then y' is obtained from y by adding $\varDelta/2$ ones on the left side, and removing $\varDelta/2$ ones from the right side. By Observation 1, $|x'|_0 = |x'|_1 = |y'|_0 = |y'|_1 = p$. □

Lemma 3. *If $k > 16$, there exists $r_k = \varOmega\big(\sqrt{|\mathbf{w}_k|}\big)$ such that for every $\varGamma \in \mathbf{S}_k$*

$$\mathrm{neigh}^+(\varGamma, r_k) \subseteq SQV(\mathbf{w}_k) \quad \textit{or} \quad \mathrm{neigh}^-(\varGamma, r_k) \subseteq SQV(\mathbf{w}_k).$$

Proof. For $\varGamma \in \mathbf{S}_k$ we define i, j, and the Abelian square $x'y'$ corresponding to $\varGamma$ as in the proof of Lemma 2. Let β and α be the distances from the right end of x' to the beginning and to the end of the block 1^{i-1}; see Fig. 2. Similarly we define δ as the distance of the right end of y' to the left endpoint of the block 1^j. One can easily check that the distance of the right end of y' to the end of the block 1^j equals α (see Fig. 2).

Note that $\alpha + \beta = i - 1 \ge \lceil \frac{3}{4}k \rceil - 1$ and $\delta \ge \beta$. There are two cases:

(a) If $\alpha \ge \beta$, then $\alpha \ge (i-1)/2$. Then $\mathrm{neigh}^+(\varGamma, \lfloor \alpha/2 \rfloor) \subseteq SQV(\mathbf{w}_k)$.
(b) If $\alpha < \beta$, then $\beta \ge (i-1)/2$. Then $\mathrm{neigh}^-(\varGamma, \lfloor \beta/2 \rfloor) \subseteq SQV(\mathbf{w}_k)$.

Thus we set $r_k = \lfloor \big(\lceil \frac{3}{4}k \rceil - 1\big)/4 \rfloor = \varOmega\big(\sqrt{|\mathbf{w}_k|}\big)$ and the conclusion holds. □

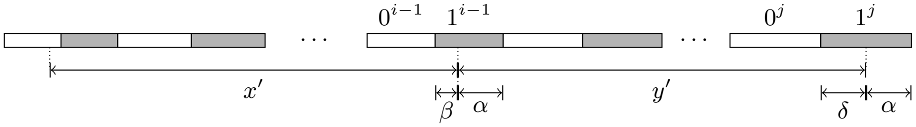

Fig. 2. Illustration of the proof of Lemma 3. Observe that the number of ones to the right of x' and to the right of y' is the same, due to the construction of Lemma 2.

Theorem 2. $SQ'_{\text{Abel}}(n) = \Omega(n^{1.5}/\log n)$.

Proof. We have constructed the family of words $\mathbf{w}_k$ together with the sets $\mathbf{S}_k$, which we show (Lemma 2) to be square vectors of $\mathbf{w}_k$. Due to Lemma 1 we have

$$|\mathbf{S}_k| = \Omega(|\mathbf{w}_k|/\log|\mathbf{w}_k|).$$

Note that for any $\Gamma_1, \Gamma_2 \in \mathbf{S}_k$, $\Gamma_1 \neq \Gamma_2$, and $r \geq 0$ we have $\text{neigh}(\Gamma_1, r) \cap \text{neigh}(\Gamma_2, r) = \emptyset$. Thus by Lemma 3 we obtain

$$SQ'_{\text{Abel}}(\mathbf{w}_k) = |SQV(\mathbf{w}_k)| \geq |\mathbf{S}_k| r_k = \Omega(|\mathbf{w}_k|^{1.5}/\log|\mathbf{w}_k|).$$

This completes the lower bound proof for $n = |\mathbf{w}_k|$. Otherwise we pick the longest word $\mathbf{w}_k$, $|\mathbf{w}_k| \leq n$, and append it with sufficiently many zeros. □

2.2 An Upper Bound for $SQ'_{\text{Abel}}(n, 2)$

The number of blocks in a word w is defined as:

$$\#_{bl}(w) = 1 + |\{1 \leq i < |w| \ : \ w[i] \neq w[i+1]\}|.$$

For example $\#_{bl}(\mathbf{w}_k) = 2k$. We show a nontrivial upper bound for the number of nonequivalent Abelian squares in words with a given number m of blocks.

Lemma 4. *For a word w and a nonnegative integer δ suppose the following subwords are uniform but not all equal:*

$$w_1 = w[j..j+\delta],\ w_2 = w[j+k..j+k+\delta],\ w_3 = w[j+2k..j+2k+\delta].$$

If $w[j..j+2k-1]$ is an Abelian square, then no Abelian square of the same length starts at any position in the interval $[j+1..j+\delta]$.

Proof. Due to the binary alphabet we have exactly three cases: $w_1 = w_3$, $w_1 = w_2$ or $w_2 = w_3$. We prove the lemma only in the first case; see Fig. 3. The remaining cases admit similar proofs.

Let $w_1 = w_3 = (c_1)^\delta$ and $w_2 = (c_2)^\delta$ with $c_1 \neq c_2$. Denote $u_1 = w[j..j+k-1]$ and $u_2 = w[j+k..j+2k-1]$. Whenever we shift u_1 to the right (by at most δ positions), the number of occurrences of c_1 decreases and the number of occurrences of c_2 increases. However, when we shift u_2 to the right, the number of occurrences of c_1 increases and the number of occurrences of c_2 decreases. Therefore, we cannot obtain an Abelian square. □

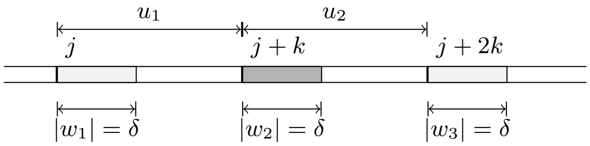

Fig. 3. Illustration of Lemma 4: the shaded areas correspond to uniform subwords (the first and the third one are composed of the same letter). An Abelian square u_1u_2 at position j excludes any Abelian square of the same length starting in the shaded area to the right of j.

Theorem 3. *If w is a binary word of length n with m blocks, then*

$$SQ'_{\text{Abel}}(w) \leq 3(m+1)\tfrac{n}{2}.$$

Proof. We call $\{0, 1, \ldots, n\}$ the set of *interpositions* of w. Intuitively, interpositions can be interpreted as locations between two consecutive letters, before the first letter or after the last letter. Let A, the set of *alternating* interpositions, contain 0, n and all interpositions i for which $w[i] \neq w[i+1]$. In other words if i is not alternating, then $w[i] = w[i+1]$, in particular both these letters are well defined. Note that $|A| = m+1$. To each Abelian square $w[i..i+2k-1]$ we assign three interpositions which we call *special*: the first interposition $i-1$ ($\mathtt{F}$), the middle interposition $i+k-1$ ($\mathtt{M}$), and the last interposition $i+2k-1$ ($\mathtt{L}$).

For each square vector $\Gamma \in SQV(w)$ we consider only the rightmost occurrence of an Abelian square corresponding to Γ.

First, we consider Abelian squares for which one of the special interpositions is alternating. Let $v = w[i..i+2k-1]$ be such an Abelian square. We uniquely label v with a triple representing an alternating interposition, the type of this interposition ($\mathtt{F}$/$\mathtt{M}$/$\mathtt{L}$) and the half of v's length: if $(i-1) \in A$, then the triple is $(i-1,\ \mathtt{F},\ |v|/2)$, otherwise if $(i+k-1) \in A$, then it is $(i+k-1,\ \mathtt{M},\ |v|/2)$, and otherwise it is $(i+2k-1,\ \mathtt{L},\ |v|/2)$.

As a second group we consider the remaining (rightmost) Abelian squares. Let $v = w[i..i+2k-1]$ be such an Abelian square. Note that $w[i] = w[i+k] = w[i+2k]$ could not hold, otherwise v would not be the rightmost occurrence (v would be Abelian equivalent to $w[i+1..i+2k]$). Let ℓ_1 be the length of the maximal prefix of $w[i..n]$ of form $w[i]^*$, likewise ℓ_2 be the length of the maximal prefix of $w[i+k..n]$ of form $w[i+k]^*$, and ℓ_3 be the length of the maximal prefix of $w[i+2k..n]$ of form $w[i+2k]^*$. Let $\ell = \min(\ell_1, \ell_2, \ell_3) > 0$. We uniquely label v with a triple representing an alternating interposition, the type of this interposition and the half of v's length: if $\ell_1 = \ell$, then the triple is $(i+\ell-1,\ \mathtt{F},\ |v|/2)$, otherwise if $\ell_2 = \ell$, then it is $(i+k+\ell-1,\ \mathtt{M},\ |v|/2)$, and otherwise it is $(i+2k+\ell-1,\ \mathtt{L},\ |v|/2)$.

Lemma 4 implies that each Abelian square receives a different label. Therefore there are at most $3(m+1)\frac{n}{2}$ Abelian squares in total. □

In particular, Theorem 3 implies the following result:

Observation 2. $SQ'_{\text{Abel}}(\mathbf{w}_k) = O(|\mathbf{w}_k|^{1.5})$.

3 Bounds for Order-Preserving Squares

Recall that uv is an order-preserving square if $|u| = |v|$ and there exists a strictly increasing bijection $f : \mathrm{Alph}(u) \to \mathrm{Alph}(v)$ such that $v[i] = f(u[i])$ for all $i = 1, \ldots, |u|$. We start with an auxiliary abstract fact in which we do not require f to be of any particular monotonicity.

Lemma 5. *Let w be a word of length n over an alphabet Σ, and let Σ_1, Σ_2 be two distinct subsets of Σ of the same cardinality. Also, let f be a given bijection between Σ_1 and Σ_2. Then there are at most n distinct subwords of w of the form $xf(x)$, where $\mathrm{Alph}(x) = \Sigma_1$.*

Proof. Suppose a word $xf(x)$, where $\mathrm{Alph}(x) = \Sigma_1$, starts at position i in w. Let $j > i$ be the first occurrence of a letter in $\Sigma_2 - \Sigma_1$. Suppose it is the letter c. This letter is located in $f(x)$. Let $k \geq i$ be the first occurrence of $f^{-1}(c)$. Then $|x| = j - k$ and this determines the word $xf(x)$ as $w[i..i + 2(j - k) - 1]$.

Consequently there is at most one occurrence of a subword of the required form starting at a given position, so the number of such distinct subwords does not exceed n. □

Theorem 4. *If $k = O(1)$, then $SQ_{\mathrm{op}}(n, k) = \Theta(n)$.*

Proof. Let w be a word of length n over a k-letter alphabet Σ. Each order-preserving square is of the form $xf(x)$ where $f : \mathrm{Alph}(x) \to \mathrm{Alph}(f(x))$ is a strictly increasing bijection. If $\mathrm{Alph}(x) = \mathrm{Alph}(f(x))$, then f must be the identity and thus $xf(x)$ is an ordinary square. However, there are at most $2n$ such squares in w [14]. Otherwise, $\mathrm{Alph}(x), \mathrm{Alph}(f(x))$ and f satisfy the assumptions of Lemma 5. The number of such triples is constant with respect to n, which, combined with Lemma 5, completes the proof. □

4 Bounds for Parameterized Squares

In this section we consider words over the binary alphabet $\{0, 1\}$. An *antisquare* is a nonempty word of the form $x\bar{x}$, where $\bar{x}$ denotes bitwise negation of x. For example, 011 100 is an antisquare. Recall that uv is a parameterized square if $|u| = |v|$ and there exists a bijection $f : \mathrm{Alph}(u) \to \mathrm{Alph}(v)$ such that $v[i] = f(u[i])$ for all $i = 1, \ldots, |u|$. Observe that for binary alphabet each parameterized square is an ordinary square or an antisquare.

We also introduce *almost-squares*, which are the words of the form xax, where x is a word and $a \in \{0, 1\}$. Equivalently, an almost-square is an ordinary square with the last letter missing. The following words are examples of almost-squares: 011 1 011, 11111 0 11111, 0.

For a binary word w of length n we define the word $\hat{w}$ of length $n - 1$ so that $\hat{w}[i] = 1$ if $w[i] = w[i + 1]$ and $\hat{w}[i] = 0$ otherwise. For example, for $w = 00110101100010$ we have $\hat{w} = 1010000101100$.

For a word x of length ℓ we construct a rooted directed labeled tree $T(x)$ as follows. We start with a single path with edges labeled with the consecutive

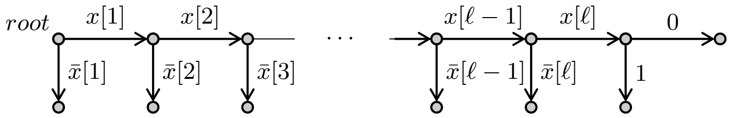

Fig. 4. Rooted directed labeled tree $T(x)$, $|x| = \ell$.

letters of x. Then we attach leaves to all nodes of the path so that each of them has two outgoing edges, one labeled with 0 and one labeled with 1; see Fig. 4. A square in a directed labeled tree is defined as a directed path such that the label of the path is an (ordinary) square.

Observation 3. *The following are equivalent:*
(a) the subword $w[i..j]$ is a parameterized square,
(b) the subword $\hat{w}[i..j-1]$ is an almost-square,
(c) the subword $\hat{w}[i..j-1]a$ for some $a \in \{0,1\}$ is a square in the tree $T(\hat{w})$.

The proof of the following lemma is an immediate generalization of the proof of the analogous upper bound on the number of ordinary squares in a word. It suffices to note that there are at most two topmost occurrences of distinct squares ending at each node of the tree; see [14,18].

Lemma 6. *In a labeled directed rooted tree with m nodes there are at most $2m$ distinct squares.*

Theorem 5. $SQ_{\text{param}}(n,2) \le 8n$.

Proof. Let w be a word of length n. Observe that $T(\hat{w})$ has at most $2n$ nodes. It follows from Observation 3 and Lemma 6 that the number of distinct almost-squares in w is at most $4n$. For each almost-square v there are exactly two parameterized squares u_1, u_2 such that $\hat{u_1} = \hat{u_2} = v$ ($\bar{u_1} = u_2$ and u_1, u_2 are both ordinary squares or both antisquares). Consequently $SQ_{\text{param}}(w)$ does not exceed twice the number of distinct almost-squares in $\hat{w}$. Hence $SQ_{\text{param}}(w) \le 8n$. □

5 Infinite Words Avoiding Nonstandard Squares/Cubes

It is known that there are infinite words over a 4-letter alphabet avoiding Abelian squares while over 3-letter alphabets such words do not exist [20]. Here, we investigate an analogous problem for other nonstandard repetitions.

We say that a word is op-square-free if it does not contain an order-preserving square of length greater than 2. Let $\Sigma_3 = \{0,1,2\}$ ordered in the natural way. Consider the morphism:

$$\psi \;\; : \; 0 \mapsto 10,\; 1 \mapsto 11,\; 2 \mapsto 12.$$

Lemma 7. *If a word $w \in \Sigma_3^*$ is square free, then $\psi(w)$ is op-square-free.*

Proof. Let $\approx$ denote the order-preserving equivalence (i.e., $u \approx v$ if $|u| = |v|$ and uv is an order-preserving square). We have the following simple observation.

Observation 4. *For any symbols $a, b, c \in \Sigma_3$ we have:*
(a) $1\,a \approx 1\,b \;\Leftrightarrow\; a = b$*;*
(b) $a\,1\,b \approx 1\,c\,1 \;\Rightarrow\; a = b$*.*

Suppose to the contrary that $w' = \psi(w)$ contains an order-preserving square $u'v' = w'[i..i{+}2k{-}1]$, with $|u'| = |v'| = k \geq 2$. We consider four cases depending on the parity of i and k.

If $2 \mid k$ and $2 \nmid i$, then u' and v' start with a 1 and every second symbol of each of them is a 1. Consequently, by Observation 4(a), $u' = v'$. Moreover, in this case we have $u' = \psi(u)$ and $v' = \psi(v)$ for some subword uv of w. Hence, uv is a square in w, a contradiction.

If $2 \mid k$ and $2 \mid i$, then $w'[i-1...i+2k-2]$ is also an order-preserving square. The conclusion follows from the previous case.

If $2 \nmid k$ and $2 \nmid i$, then u' and v' start with $1c1$ and $a1b$ for some $a, b, c \in \Sigma_3$, respectively. By Observation 4(b) we conclude that $a = b$, which implies a square ab in w, a contradiction.

The final case, $2 \nmid k$ and $2 \mid i$, also implies a 2-letter square in w just as in the previous case. This completes the proof that w' is op-square-free. □

We apply Lemma 7 to all prefixes of an infinite square-free word [24] over a ternary alphabet and obtain the following result.

Theorem 6. *There exists an infinite op-square-free word over 3-letter alphabet.*

A parameterized cube is a word uvw such that both uv and vw are parameterized squares. A word is called parameterized-square-free (parameterized-cube-free) if it does not contain parameterized squares (parameterized cubes) of length greater than 3. We show that there is no infinite parameterized-square-free word and construct a binary parameterized-cube-free word.

Theorem 7. *There is no infinite parameterized-square-free word.*

Proof. Suppose to the contrary that such an infinite word x exists. In the proof we denote symbols of $\mathrm{Alph}(x)$ by a, b, c, d. Note that every suffix of x has to contain two adjacent equal symbols. This is because $abcd$ for $a \neq b$ and $c \neq d$ is a parameterized square. Moreover, x has to contain some three adjacent equal symbols. The reason is that $abbd$ for $a \neq b \neq d$ is a parameterized square.

We can therefore assume that x contains a subword aaa. To avoid a parameterized square of length 4, this subword must be followed in x by some letter $b \neq a$. For the same reason the next letter c must satisfy $c \neq b$, and afterwards the subword $aaabc$ must be followed by two more occurrences of c. Finally the next letter must be $d \neq c$ to avoid a parameterized square $cccc$. We conclude that x contains a subword $aaabcccd$ for $b \neq a$ and $d \neq c$, which turns out to be a parameterized square. This contradiction completes the proof. □

Let τ be the infinite Thue-Morse word. Recall that τ is cube-free [25]. Also recall the morphism ψ defined just before Lemma 7.

Theorem 8. *The word $\psi(\tau)$ is parameterized-cube-free.*

Proof. Suppose to the contrary that $u_1u_2u_3$ is a parameterized cube in $\psi(\tau)$, with $|u_1| = |u_2| = |u_3| = k > 1$. Note that $\psi(\tau)$ does not contain 6 ones in a row. Hence, at least one of the words u_1, u_2, u_3 contains 0, therefore each of them contains 0. Moreover every second symbol of u_1, u_2, u_3 is 1.

Recall from Section 4 that a binary parameterized square is either an ordinary square or an antisquare. If $2 \mid k$, then the ones of every second position of u_1, u_2, u_3 align and u_1u_2, u_2u_3 must be ordinary squares. Therefore $u_1u_2u_3$ is an ordinary cube in $\psi(\tau)$ which induces a cube in τ.

If $2 \nmid k$, the same argument implies that both u_1u_2 and u_2u_3 are antisquares. Because of the ones on every second position of u_1, u_2, u_3 we actually have $u_1 = 0101\cdots$, $u_2 = 1010\cdots$, $u_3 = 0101\cdots$ or $u_1 = 1010\cdots$, $u_2 = 0101\cdots$, $u_3 = 1010\cdots$. In both cases we obtain a cube $(10)^3$ in $\psi(\tau)$ which induces 0^3 in τ. □

6 Final Remarks

We have presented several combinatorial results related to the maximum number of nonstandard squares in a word of length n. For Abelian squares we have shown that $SQ_{\text{Abel}}(n,2) = \Theta(n^2)$ and $SQ'_{\text{Abel}}(n,2) = \Omega(n^{1.5}/\log n)$. The latter bound, although reached by a simple family of words, required a rather involved proof.

For squares in order-preserving and parameterized setting we show that their maximum number is linear of n for a constant and a binary alphabet, respectively. We have also presented examples of infinite words over a minimal alphabet that avoid squares in order-preserving setting and cubes in parameterized setting, respectively.

The main open question that arises from our work is to provide an upper bound for $SQ'_{\text{Abel}}(n,2)$. We have made a step towards this bound by showing that the maximum number of distinct Abelian squares in a word of length n containing m blocks is $O(nm)$. The remaining open questions are connected to $SQ'_{\text{op}}(n,k)$ and $SQ'_{\text{param}}(n,k)$ for arbitrary k (not necessarily a constant). Based on experimental results, we state the following conjecture:

Conjecture 1. *$SQ'_{\text{Abel}}(n,2) = O(n^{1.5})$, $SQ'_{\text{op}}(n,k) = SQ'_{\text{param}}(n,k) = \Theta(n)$ for any $k \geq 2$.*

References

1. Baker, B.S.: Parameterized pattern matching: Algorithms and applications. Journal of Computer and System Sciences 52(1), 28–42 (1996)
2. Berstel, J., Karhumäki, J.: Combinatorics on words: a tutorial. Bulletin of the EATCS 79, 178–228 (2003)
3. Blanchet-Sadri, F., Jiao, Y., Machacek, J.M., Quigley, J., Zhang, X.: Squares in partial words. Theoretical Computer Science 530, 42–57 (2014)
4. Blanchet-Sadri, F., Mercaş, R., Scott, G.: Counting distinct squares in partial words. Acta Cybernetica 19(2), 465–477 (2009)

5. Crochemore, M., Ilie, L., Rytter, W.: Repetitions in strings: Algorithms and combinatorics. Theoretical Computer Science 410(50), 5227–5235 (2009)
6. Crochemore, M., Iliopoulos, C.S., Kociumaka, T., Kubica, M., Langiu, A., Pissis, S.P., Radoszewski, J., Rytter, W., Waleń, T.: Order-preserving incomplete suffix trees and order-preserving indexes. In: Kurland, O., Lewenstein, M., Porat, E. (eds.) SPIRE. LNCS, vol. 8214, pp. 84–95. Springer (2013)
7. Crochemore, M., Iliopoulos, C.S., Kociumaka, T., Kubica, M., Radoszewski, J., Rytter, W., Tyczyński, W., Waleń, T.: The maximum number of squares in a tree. In: Kärkkäinen, J., Stoye, J. (eds.) CPM. LNCS, vol. 7354, pp. 27–40. Springer (2012)
8. Crochemore, M., Iliopoulos, C.S., Kubica, M., Radoszewski, J., Rytter, W., Waleń, T.: Extracting powers and periods in a word from its runs structure. Theoretical Computer Science 521, 29–41 (2014)
9. Erdős, P.: An asymptotic inequality in the theory of numbers (in Russian). Vestnik Leningrad University: Mathematics 15, 41–49 (1960)
10. Erdős, P.: Some unsolved problems. Hungarian Academy of Sciences Mat. Kutató Intézet Közl. 6, 221–254 (1961)
11. Evdokimov, A.A.: Strongly asymmetric sequences generated by a finite number of symbols. Doklady Akademii Nauk SSSR 179(6), 1268–1271 (1968)
12. Fici, G.: Personal communication
13. Fici, G., Langiu, A., Lecroq, T., Lefebvre, A., Mignosi, F., Prieur-Gaston, É.: Abelian repetitions in Sturmian words. In: Béal, M.P., Carton, O. (eds.) DLT 2013. LNCS, vol. 7907, pp. 227–238. Springer (2013)
14. Fraenkel, A.S., Simpson, J.: How many squares can a string contain? Journal of Combinatorial Theory, Series A 82, 112–120 (1998)
15. Gusfield, D., Stoye, J.: Linear time algorithms for finding and representing all the tandem repeats in a string. Journal of Computer and System Sciences 69(4), 525–546 (2004)
16. Hardy, G., Wright, E., Heath-Brown, D., Silverman, J.: An Introduction to the Theory of Numbers. Oxford mathematics, OUP Oxford (2008)
17. Huova, M., Karhumäki, J., Saarela, A.: Problems in between words and abelian words: k-abelian avoidability. Theor. Comput. Sci. 454, 172–177 (2012)
18. Ilie, L.: A simple proof that a word of length n has at most $2n$ distinct squares. Journal of Combinatorial Theory, Series A 112(1), 163–164 (2005)
19. Ilie, L.: A note on the number of squares in a word. Theoretical Computer Science 380(3), 373–376 (2007)
20. Keränen, V.: Abelian squares are avoidable on 4 letters. In: Kuich, W. (ed.) ICALP. LNCS, vol. 623, pp. 41–52. Springer (1992)
21. Muthukrishnan, S.: New results and open problems related to non-standard stringology. In: Galil, Z., Ukkonen, E. (eds.) Combinatorial Pattern Matching, LNCS, vol. 937, pp. 298–317. Springer Berlin Heidelberg (1995)
22. Muthukrishnan, S., Palem, K.: Non-standard stringology: Algorithms and complexity. In: Proceedings of the Twenty-sixth Annual ACM Symposium on Theory of Computing. pp. 770–779. STOC '94, ACM, New York, NY, USA (1994)
23. Pleasants, P.A.: Non-repetitive sequences. Mathematical Proceedings of the Cambridge Philosophical Society 68, 267–274 (1970)
24. Thue, A.: Über unendliche Zeichenreihen. Norske Videnskabers Selskabs Skrifter Mat.-Nat. Kl. 7, 1–22 (1906)
25. Thue, A.: Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. Norske Videnskabers Selskabs Skrifter Mat.-Nat. Kl. 10, 1–67 (1912)