A Stronger Square Conjecture on Binary Words

Nataša Jonoska[1], Florin Manea[2], and Shinnosuke Seki[3,4]

[1] Department of Mathematics and Statistics, University of South Florida, 4202 East Fowler Avenue, Tampa, FL 33620, USA
jonoska@math.usf.edu
[2] Institut für Informatik, Christian-Albrechts-Universität zu Kiel, Kiel, Germany
flm@informatik.uni-kiel.de
[3] Helsinki Institute for Information Technology (HIIT)
[4] Department of Information and Computer Science, Aalto University, P. O. Box 15400, FI-00076, Aalto, Finland
shinnosuke.seki@aalto.fi

Abstract. We propose a stronger conjecture regarding the number of distinct squares in a binary word. Fraenkel and Simpson conjectured in 1998 that the number of distinct squares in a word is upper bounded by the length of the word. Here, we conjecture that in the case of a word of length n over the alphabet $\{a, b\}$, the number of distinct squares is upper bounded by $\frac{2k-1}{2k+2}n$, where k is the least of the number of a's and the number of b's. We support the conjecture by showing its validity for several classes of binary words. We also prove that the bound is tight.

1 Conjectures

Let Σ be an alphabet and Σ^* be the set of all words over Σ. Let $w \in \Sigma^*$. By $|w|$, we denote its length. For a letter $a \in \Sigma$, we denote the number of occurrences of a's in w by $|w|_a$. In this paper, n exclusively denotes the length of a word in which squares are to be counted.

Let $\mathrm{Sq}(w) = \{uu \mid w = xuuy \text{ for some } x, y \in \Sigma^* \text{ with } w \neq xy\}$ be the set of all squares occurring in w. Its size, denoted by $\#\mathrm{Sq}(w)$, has been conjectured to be bounded from above by the length of w [1].

That is to say, $\#\mathrm{Sq}(w) \leq n$ for any word w of length n; a slightly stronger conjecture is $\#\mathrm{Sq}(w) \leq n - |\Sigma|$, given in [2]. Notable upper bounds shown so far are $\#\mathrm{Sq}(w) \leq 2n$ [1], further improved by Ilie to $\#\mathrm{Sq}(w) \leq 2n - \log n$ [3], this being the best bound known so far.

An infinite word, over the binary alphabet $\Sigma_2 = \{a, b\}$, whose finite factors have a relatively large number of distinct squares compared to their length was given by Fraenkel and Simpson [1]:

$$w_{\mathrm{fs}} = a^1ba^2ba^3ba^2ba^3ba^4ba^3ba^4ba^5ba^4ba^5ba^6b\cdots . \tag{1}$$

None of its factors of length n with k letters b contain more than $\frac{2k-1}{2k+2}n$ distinct squares (Corollary 2). In fact, we propose (Conjecture 1) that this upper bound holds not only for the factors of w_{fs} but for all binary words. A computer

V. Geffert et al. (Eds.): SOFSEM 2014, LNCS 8327, pp. 339–350, 2014.

program verified the conjecture for all binary words of length less than 30, as well as for randomly generated binary words of length up to 500 without any counterexample found. Due to this, we propose, as our first contribution, the following stronger conjecture regarding the number of squares.

Conjecture 1. Let $k \geq 2$. For any binary word $w \in \Sigma_2^+$ of length n with k b's where $k \leq \lfloor \frac{n}{2} \rfloor$,

$$\#\mathrm{Sq}(w) \leq \frac{2k-1}{2k+2} n.$$

The bound is defined here as a function of the number of b's. However, Conjecture 1 gives an upper bound on number of squares more generally by redefining k as $\min\{|w|_a, |w|_b\}$, as the number of distinct squares in a binary word is invariant under the isomorphism swapping the letters a and b. Another conjecture proposed in [2] states that for a binary word w we have $\#\mathrm{Sq}(w) \leq n - 2$. Our conjecture is, however, stronger, because $\frac{2k-1}{2k+2} n \leq n - 2$ whenever $2 \geq k \leq \lfloor \frac{n}{2} \rfloor$.

Conjecture 1 doesn't consider words with at most one b because they are square sparse. It is clear that $\mathrm{Sq}(a^n) = \{(a)^2, (aa)^2, \ldots, (a^{\lfloor n/2 \rfloor})^2\}$, and hence, $\#\mathrm{Sq}(a^n) = \lfloor n/2 \rfloor$. The sole b in a word cannot be a part of any square, so its presence cannot increase the number of squares. Thus, the upper bound $\lfloor n/2 \rfloor$ holds canonically for any binary word with at most one b.

Note that our conjecture not only strengthens the conjecture that $\#\mathrm{Sq}(w) \leq n$, but its dependency on the number of b's suggests that a possible proof might be obtained by induction on this number. We show here that it holds when at most nine b's are present in the word.

Parenthesizing the sequence of positive integers representing the powers of a's in the word w_{fs}, in a convenient manner gives the sequence $(1, 2, 3)$, $(2, 3, 4)$, $(3, 4, 5), \ldots$. This reveals the structure of w_{fs} as catenation of simpler words $a^i b a^{i+1} b a^{i+2} b$, $i = 1, 2, \ldots$. As another contribution, we propose a structurally simpler infinite word, whose coefficients just increment:

$$w_{\mathrm{jms}} = a^1 b a^2 b a^3 b a^4 b a^5 b a^6 b \cdots, \tag{2}$$

and prove that it is quite rich with respect to the number of squares its factors contain. Indeed, we show that its factors achieve the upper bound in Conjecture 1 asymptotically.

The word w_{jms} points out that a word does not necessarily need a complicated structure in order to have many squares. Thus, we further prove that for any word w of length n with k letters b, whose coefficient sequence is sorted (incrementing or decrementing), Conjecture 1 holds (see Theorem 2). This result follows by induction on the number of b's on the word.

As an important technical tool, our analysis is not based on combinatorial properties that the word itself has, but rather on the combinatorial properties of the sequence of powers of the letters a (called here "coefficient sequence"). This allows us to define more general classes of words for which the conjecture holds (e.g., Theorem 3).

2 Preliminaries

Let Σ be an alphabet; for this section this alphabet can even be infinite (for instance, the set of positive integers). For words $u, v \in \Sigma^*$, v is a *prefix* (*suffix*) of u if $u = vy$ (resp. $u = xv$) for some word $y \in \Sigma^*$ (resp. $x \in \Sigma^*$). If $u \neq v$, v is called a *proper prefix* (resp. *proper suffix*). The prefix and proper prefix relations are denoted by $v \leq_p u$ and $v <_p u$, respectively. The suffix and proper suffix relations, $\geq_s$ and $>_s$, are defined analogously. If $u = xvy$, then v is a *factor* of u. A factor that is not a prefix or suffix is said to be *proper*.

Three square lemmas concern the occurrence of two squares at the same location in the word, with another square there, or "nearby" (see, e.g., [4,5,3,6]). We give an analogous lemma, not on squares, but on words of the form uau as Lemma 2, which plays an important part in our inductive analysis. Its proof is a modification of the proof of Theorem 1 in [5], but based on the variant of synchronization lemma below.

Lemma 1. *Let $x, y \in \Sigma^*$ and $a \in \Sigma$ be such that xay is primitive of length at least 2. If $(xay)^2 = z_1yxz_2$ for some $z_1, z_2 \in \Sigma^*$, then $z_1 = xa$ and $z_2 = ay$.*

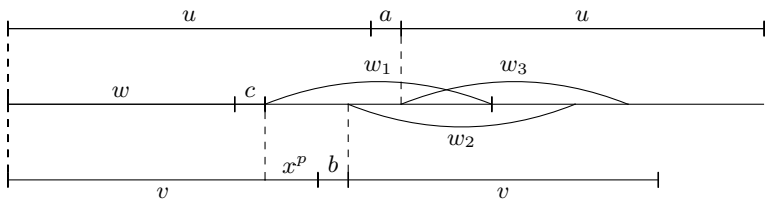

Fig. 1. Three words uau, vbv, wcw starting at the same position

Lemma 2. *For words $u, v, w \in \Sigma^*$ and letters $a, b, c \in \Sigma$, if $u \leq_p wcw <_p vbv <_p uau$ holds, then the word $cwcwc$ occurs as a factor on vbv.*

Proof. Note that the prefix relations imply $|w| \geq 1$ (hence, $|v| \geq 2$ and $|u| \geq 3$). As illustrated in Fig. 1, we denote the second occurrence of w in wcw by w_1, the prefix w of the second v by w_2, and the prefix w of the second u by w_3.

Let $v = wcx^p$ for some primitive word x and $p \geq 0$. If $p = 0$, then $w = b^i$ and $v = b^ic$ for some $i \geq 1$. Note that $|ua| < |vbv| < |ua| + |v|$ and $v = b^ic <_p u$. These mean that the rightmost letter c of the second v is on the prefix b^i of the second u, and hence, $b = c$. Now we have $w = c^i$ and $vbv = c^{2i+3} = cwcwc$.

The case of p being positive is considered below. Note that w_2 certainly overlaps with w_3, while it does not overlap with w_1 if and only if $u = wcw = vb$ holds. We handle the non-overlapping subcase first. In this subcase, $u = vb$ implies $v = a^j$ and $u = a^jb$ for some $j \geq 2$. With $u = wcw$, they give $c = a$ and

w is a power of a, and hence, $a = b = c$. Thus, all of wcw, vbv, uau are a power of a. Now we examine the other case when w_1 and w_2 overlap. Let $x^p b = y^s$ for some primitive word y and $s \geq 1$. The overlap gives $w = y^q y'$ for some $q \geq s$, where y' is a proper prefix of y such that $y = y'y''$ for some $y'' \in \Sigma^+$. If $|y| = 1$, then $w = a^q$, $v = a^q c a^p$, and $p < s \leq q$. Since w_3 is a proper factor of v, we can conclude that $c = a$. If $|y| \geq 2$, on the other hand, we can apply Lemma 1 to the overlap between w_2 and w_3, which is of length at least $|y| - 1$, and obtain that the overlap is $y^r y'$ for some $s \leq r \leq q$. The remaining suffix of w_3, which is $y''y^{q-r-1}y'$, and $(w^2)^{-1}v = cy^s b^{-1}$ begin at the same position. Then we have $y''y' = cyb^{-1}$, and this means $b = c$ because $y''y'$ is a conjugate of y. Now synchronization gives us $y'' = c$ and $y' = yc^{-1}$. Then $w = y^{q+1}c^{-1}$, and hence, $wcw = y^{2(q+1)}c^{-1}$. We also have $v = wbx^p = y^{q+1+s}c^{-1}$, and this gives $vbv = y^{2(s-1)}yc^{-1}cy^{2(q+1)}c^{-1}cyc^{-1} = y^{2(s-1)}yc^{-1}\underline{cwcwc}\, yc^{-1}$. □

The results of this section will not be applied for binary words, in which we count squares, but rather for their coefficient sequences, which are words over integer alphabets. These two lemmas enable us to develop a series of technical tools, which are important to our analysis, e.g., Lemmas 6 and 7.

3 Counting Squares

In this section we show that the bound in Conjecture 1 is tight and factors of w_{jms} with k b's achieve it. Throughout the paper, we denote the binary word with k b's (and at least k a's) in which we count squares by $w_k = a^{i_0}ba^{i_1}b\cdots ba^{i_k}$, where $i_0, \ldots, i_k \geq 0$, and assume that it is of length n. We represent w_k simply as $\langle i_0, \ldots, i_k\rangle$ called the *coefficient sequence* of w_k. We define the *coefficient set* of w_k to be the multiset $I(w_k) = \{i_0, i_1, \ldots, i_k\}$. The cardinality of $I(w_k)$ is considered to be $k+1$ and is denoted $|I(w_k)|$. The argument w_k is omitted from $I(w_k)$ when it is understood in the context. For $j \leq k+1$, by $I[j]$ we denote the *j-th smallest element* of I. Since its maximum element $I[k+1]$ is often referred to, it is more convenient to denote it by $I[\max]$. More generally, by $I[\max -(j-1)]$ we mean the j-th largest element of I.

Squares in w_k that are free from b can be counted simply by checking the largest coefficient in $I(w_k)$ as:

$$\#(\mathrm{Sq}(w_k) \cap a^+) = \left\lfloor \frac{I(w_k)[\max]}{2} \right\rfloor. \tag{3}$$

In counting squares including b's, we first classify them with respect to the equivalence relation "cyclic shift of a's", and then do counting per class. For instance, $x = a^3baba^3bab$ and $ababa^3baba^2$ are members of the same equivalence class because cyclically shifting the first two a's transforms the former into the latter. The class of x contains other two words a^2baba^3baba and $baba^3baba^3$. In contrast, aba^3baba^3b does not belong to the same class as x because one has to shift a b in order to obtain this word from x.

In general, a binary square uu with $2m$ b's (m b's per u) has a coefficient sequence $\langle i_0, i_1, \ldots, i_{2m}\rangle$ ($i_0, \ldots, i_{2m} \geq 0$) such that $i_0 + i_{2m} = i_m$ and $i_j = i_{m+j}$

for $1 \le j \le m-1$. Let $c = i_m$. The first property lets $i_{2m} = c - i_0$ and by replacing the sequences $i_1, \ldots, i_{m-1}$ and $i_{m+1}, \ldots, i_{2m-1}$ by μ, the square can be written as $\langle i_0, \mu, c, \mu, c - i_0 \rangle$. The squares that result from applying cyclic shift of a's to uu are those written as $\langle c, \mu, c, \mu, 0 \rangle$, $\langle c-1, \mu, c, \mu, 1 \rangle$, $\langle c-2, \mu, c, \mu, 2 \rangle$, $\ldots$, $\langle 0, \mu, c, \mu, c \rangle$, and they compose one equivalence class. We denote the equivalence class with $\langle \mu, c, \mu \rangle$. Its cardinality is $c + 1$.

By $\#\mathrm{Sq}_{\langle \mu,c,\mu \rangle}(w)$, we denote the number of squares in the class $\langle \mu, c, \mu \rangle$ that occur in a binary word w. Clearly, $\#\mathrm{Sq}_{\langle \mu,c,\mu \rangle}(w) \le c + 1$. When the equality holds, we say that the class $\langle \mu, c, \mu \rangle$ is *saturated in* w.

Example 1. The class $\langle 1, 3, 1 \rangle$ consists of the squares a^3baba^3bab, a^2baba^3baba, $ababa^3baba^2$, and $baba^3baba^3$. It is not saturated in $a^2baba^3baba^3$ as the first square is missing, while it is saturated in $a^3baba^3baba^3$ or in $a^2baba^3baba^3bab$.

Now, we count the squares in the class $\langle \mu, c, \mu \rangle$ of the word w_k. First of all, the coefficient sequence $\langle i_0, \ldots, i_k \rangle$ of w_k must contain the class identifier $\langle \mu, c, \mu \rangle$ as its *proper* factor in order for such a square to occur in w_k. When $\langle \mu, c, \mu \rangle$ occurs exactly once, that is, there are unique coefficients $\ell, r \ge 0$ such that $\langle \ell, \mu, c, \mu, r \rangle$ is a factor of the coefficient sequence, the count is

$$\begin{aligned} \#\mathrm{Sq}_{\langle \mu,c,\mu \rangle}(w_k) &= \begin{cases} \min\{\ell, c\} + \min\{c, r\} - c + 1 & \text{if } \ell + r \ge c \\ 0 & \text{otherwise} \end{cases} \\ &\le \min\{\ell, c, r\} + 1 \end{aligned} \tag{4}$$

where 4 follows from $\min\{i, k\} + \min\{k, j\} - k \le \min\{i, j, k\}$ for $i, j, k \ge 0$.

It must be noted that (4) does not depend on μ.

We verify Conjecture 1 for a word $w_2 = a^{i_0}ba^{i_1}ba^{i_2}$. The sole class whose square can occur in w_2 is $\langle i_1 \rangle$. Using (4), squares in this class are counted in w_k as $\#\mathrm{Sq}_{\langle i_1 \rangle} \le \min\{i_0, i_1, i_2\} + 1 = I(w_2)[\max - 2] + 1$. Summing this and (3) gives

$$\begin{aligned} \#\mathrm{Sq}(w_2) &\le \frac{I(w_2)[\max]}{2} + I(w_2)[\max - 2] + 1 \\ &\le \frac{I(w_2)[\max]}{2} + \frac{1}{2}\left(n - 2 - I(w_2)[\max]\right) + 1 = \frac{1}{2}n. \end{aligned}$$

Proposition 1. *$\#\mathrm{Sq}(w_2) \le \frac{1}{2}n$ for any binary word w_2 of length n with 2 b's.*

Double-counting is a significant issue. When the coefficient sequence of a word w_k includes the factor $\langle \mu, c, \mu \rangle$ exactly twice as $\langle u \rangle = \langle \ell_1, \mu, c, \mu, r_1 \rangle$ and $\langle v \rangle = \langle \ell_2, \mu, c, \mu, r_2 \rangle$, we have

$$\begin{aligned} \#\mathrm{Sq}_{\langle \mu,c,\mu \rangle}(w_k) = {} & \#\mathrm{Sq}_{\langle \mu,c,\mu \rangle}(u) + \#\mathrm{Sq}_{\langle \mu,c,\mu \rangle}(v) \\ & - \max\big\{\min\{\ell_1, \ell_2, c\} + \min\{c, r_1, r_2\} - c + 1, 0\big\}. \end{aligned} \tag{5}$$

The subtracted term accounts for double-counting. It is 0 (i.e., u does not share any square in the class $\langle \mu, c, \mu \rangle$ with v) if and only if $\min\{\ell_1, \ell_2\} + \min\{r_1, r_2\} < c$.

Before proceeding, we note that Lemmas 1 and 2 deal with words of the form $\mu c \mu$, so, by extension, with generative classes $\langle \mu, c, \mu \rangle$. The two lemmas offer a better understanding of the combinatorial properties of such generative classes, and provide the fundamentals needed to use them in our proofs.

The tightness of the bound $\frac{2k-1}{2k+2}n$ for $k \geq 2$ follows by considering the factors of w_{jms}, defined in (2) . We parameterize by m the largest factor of w_{jms} with k b's as $w_{\text{jms},k}(m) = a^m b a^{m+1} b \cdots b a^{m+k}$. As the coefficients of such a factor are pairwise distinct, squares in any class $\langle \mu, c, \mu \rangle$ with μ being nonempty do not occur in the factor. In fact, the classes whose squares are capable of occurring are $\langle m+1 \rangle, \langle m+2 \rangle, \ldots, \langle m+k-1 \rangle$ (not being a proper factor, neither $\langle m \rangle$ nor $\langle m+k \rangle$ can find its square in $w_{\text{jms},k}$). Moreover, $w_{\text{jms},k}(m)$ contains all but one squares of each class. Due to (3) and (4), we have

$$\#\text{Sq}(w_{\text{jms},k}(m)) = \left\lfloor \frac{m+k}{2} \right\rfloor + \sum_{i=1}^{k-1} (m+i) = \left\lfloor \frac{m+k}{2} \right\rfloor + \frac{(k-1)(2m+k)}{2}.$$

Removing the floor function gives two formulae sandwiching $\#\text{Sq}(w_{\text{jms},k}(m))$ and dividing them by $|w_{\text{jms},k}(m)| = \frac{m(2k+2)+k(k+3)}{2}$ yields

$$\frac{m(2k-1)+k^2-2}{m(2k+2)+k(k+3)} \leq \frac{\#\text{Sq}(w_{\text{jms},k}(m))}{|w_{\text{jms},k}(m)|} \leq \frac{m(2k-1)+k^2}{m(2k+2)+k(k+3)}.$$

The sandwiching functions are monotonically-increasing in m for $k \geq 2$ and their limit as m approaches infinity is $(2k-1)/(2k+2)$.

4 Towards the Inductive Proof

The main aim of this section is to propose an inductive approach to Conjecture 1. The next inequality is of principal significance for this purpose.

Lemma 3. *The inequality*

$$\#\text{Sq}(w_k) \leq \left\lfloor \frac{I(w_k)[\max]}{2} \right\rfloor + \sum_{j=1}^{|I(w_k)|-2} (I(w_k)[j]+1) \tag{6}$$

implies $\#\text{Sq}(w_k) \leq \frac{2k-1}{2k+2}n$.

Proof. The inequality (6) is expanded as:

$$\begin{aligned}\#\text{Sq}(w_k) &\leq \left\lfloor \frac{I[\max]}{2} \right\rfloor + \left\lfloor \frac{k-1}{k}(n - I[\max]) \right\rfloor \leq \left\lfloor \frac{k-1}{k}n - \frac{k-2}{2k} \left\lceil \frac{n-k}{k+1} \right\rceil \right\rfloor \\ &= \left\lfloor \frac{2k-1}{2k+2}n + \frac{k-2}{2k+2} \right\rfloor = \left\lceil \frac{2k-1}{2k+2}n + \frac{k-1}{2k+2} \right\rceil - 1,\end{aligned}$$

where the first inequality follows from the fact that each term in the sum in (6) is at most $\lfloor (n - I[\max])/k \rfloor$. The second inequality is due to $I[\max] \geq \lceil (n-k)/(k+1) \rceil$, and at the end, we employ a standard conversion of floors to ceilings. Since $\#\text{Sq}(w_k)$ is an integer, this implies $\#\text{Sq}(w_k) \leq \frac{2k-1}{2k+2}n$. □

Lemma 4. *Let* $I = \{i_0, i_1, \ldots, i_{k-1}\}$ *and* $J = I \cup \{i_k\}$ *be multisets. Then* $\sum_{j=1}^{|I|-2} I[j] + \min\{J[\max -2], i_k\} = \sum_{\ell=1}^{|J|-2} J[\ell]$.

With the case of $k = 2$ as the basis (Proposition 1), induction proceeds as: choose an operation that yields w_k from another word w' with less b's. Use the induction hypothesis that w' satisfies (6) to prove that the operation does not create too many squares, and conclude that (6) holds for w_k.

One such operation is catenation. Catenating ba^{i_k} to the end of $w_{k-1} = a^{i_0}b\cdots ba^{i_{k-1}}$ yields w_k. By saying that the catenation *creates* a square, we mean that the square does not occur in w_{k-1} but occurs in w_k. Let $\langle \mu, c, \mu \rangle$ be a class of squares. In order for the catenation to create a square in this class, $\langle \mu, c, \mu \rangle$ must be a *proper suffix* of the coefficient sequence $\langle i_0, \ldots, i_{k-1} \rangle$ of w_{k-1}. When $\langle i_0, \ldots, i_{k-1} \rangle$ contains a proper suffix $\langle \mu, c, \mu \rangle$ which creates new squares we say that the class $\langle \mu, c, \mu \rangle$ is *generative* for the catenation. Observe that any saturated class in w_{k-1} cannot be generative.

4.1 Induction Based on Catenation with Single Generative Class

We further show how induction would work by verification of Conjecture 1 for words whose all coefficients, but the leftmost and rightmost, are pairwise-distinct. Proposition 1 allows us to only consider the induction step ($k \geq 3$). Let $w_{k-1} = a^{i_0}b\cdots ba^{i_{k-1}}$, and we assume (6) holds for it as an induction hypothesis. The catenation of ba^{i_k} to w_{k-1} yields $w_k = w_{k-1}ba^{i_k}$. The pairwise-distinctiveness of the coefficient sequence makes $\langle i_{k-1} \rangle$ the sole generative class for the catenation, and $\min\{i_{k-2}, i_{k-1}, i_k\} + 1$ squares are thus created due to (4). With $\min\{i_{k-2}, i_{k-1}, i_k\} \leq I(w_k)[\max -2]$, Lemma 4 verifies (6) for w_k as follows:

$$\begin{aligned}
\#\mathrm{Sq}(w_k) &\leq \left\lfloor \frac{\max\{I(w_{k-1})[\max], i_k\}}{2} \right\rfloor \\
&\quad + \sum_{j=1}^{|I(w_{k-1})|-2} (I(w_{k-1})[j] + 1) + (\min\{I(w_k)[\max -2], i_k\} + 1) \\
&\leq \frac{I(w_k)[\max]}{2} + \sum_{j=1}^{|I(w_k)|-2} (I(w_k)[j] + 1).
\end{aligned}$$

Theorem 1. *For* $k \geq 2$, *let* $w_k = a^{i_0}ba^{i_1}b\cdots ba^{i_{k-1}}ba^{i_k}$ *be a binary word of length* n *with* k *b's. If* $i_1, \ldots, i_{k-1}$ *are pairwise-distinct, then* $\#\mathrm{Sq}(w_k) \leq \frac{2k-1}{2k+2}n$.

Corollary 1. *For any factor* w_k *of* w_{jms} *with* k *b's,* $\#\mathrm{Sq}(w_k) \leq \frac{2k-1}{2k+2}n$, *where* n *is the length of* w_k.

Before proceeding to the analysis of multiple generative classes, let us introduce and examine a class of words for which the bound can be verified inductively based on catenation with single generative class. A word $w_k = a^{i_0}ba^{i_1}b\cdots ba^{i_{k-1}}ba^{i_k}$ is an *ascending (descending) slope* if $i_0 \leq i_1 \leq \cdots \leq i_k$

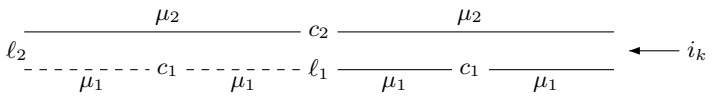

Fig. 2. Two classes $\langle \mu_2, c_2, \mu_2 \rangle, \langle \mu_1, c_1, \mu_1 \rangle$, which occur as suffixes, and hence, can be generative for the catenation of ba^{i_k} at the end

(resp. $i_0 \geq i_1 \geq \cdots \geq i_k$) holds. This notion is generalized as: w_k is a *padded slope* if its factor $a^{i_1}b \cdots ba^{i_{k-1}}$ is a slope.

Theorem 2. *For $k \geq 2$, if a binary word w_k of length n with k b's is a padded slope, then $\#\mathrm{Sq}(w_k) \leq \frac{2k-1}{2k+2}n$.*

Proof. Let $w_k = a^{i_0}ba^{i_1}b \cdots ba^{i_{k-1}}ba^{i_k}$. As induction hypothesis, assume that $w_{k-1} = a^{i_0}b \cdots ba^{i_{k-1}}$ fulfills inequality (6). Invariance of the number of squares under reversal allows us to proceed with the assumption that w_k is ascending.

Let $\langle \mu, c, \mu \rangle$ be a generative class for the catenation of ba^{i_k} to w_{k-1}. Let i_m be such that $m \geq 1$ and $i_{m-1} < i_m = \cdots = i_{k-1}$. Due to the ascending property, $\langle i_m, \ldots, i_{k-1} \rangle \geq_s \langle \mu, c, \mu \rangle$ must hold. Let $d = i_m = \cdots = i_{k-1}$. The sole class that can be generative for the catenation is $\langle d^{\lfloor (k-m)/2 \rfloor}, d, d^{\lfloor (k-m)/2 \rfloor} \rangle$ because for any $j < \lfloor (k-m)/2 \rfloor$, the class $\langle d^j, d, d^j \rangle$ has been already saturated in w_{k-1}. At most $\min\{i_\ell, d, i_k\} + 1$ squares in this class can be created due to (4), where $\ell = k - 2\lfloor (k-m)/2 \rfloor - 2$. This is clearly at most $\min\{I(w_k)[\max -2], i_k\} + 1$. Lemma 4 now concludes that the inequality (6) holds for w_k. □

Remark. We note that the proofs of Theorems 1 and 2 can be adjusted such that whenever the catenation of ba^{i_k} to w_{k-1} yields a single generative class, the Conjecture 1 holds.

4.2 Induction Based on Catenation with Multiple Generative Classes

However, catenation may involve more than one generative class. For instance, in extending the prefix $a^1ba^2ba^3ba^2$ of w_{fs} (see (1)) by ba^3, squares in the two classes $\langle 2 \rangle$ and $\langle 2, 3, 2 \rangle$ are created.

We begin with examining catenation with two generative classes. Consider two generative classes $\langle \mu_1, c_1, \mu_1 \rangle, \langle \mu_2, c_2, \mu_2 \rangle$ for the catenation of ba^{i_k} to $w_{k-1} = a^{i_0}ba^{i_1}b \cdots ba^{i_{k-1}}$, which yields $w_k = w_{k-1}ba^{i_k}$. From (4), we get that the catenation creates at most $\min\{I(w_k)[\max -3], i_k\} + \min\{I(w_k)[\max -2], i_k\} + 2$ squares in these classes. When $\langle \mu_2 \rangle \geq_s \langle \mu_1, c_1, \mu_1 \rangle$, the number turns out to be bounded by $\min\{I(w_k)[\max -2], i_k\} + 1$, as shown in the next lemma.

Lemma 5. *Let $\langle \mu_1, c_1, \mu_1 \rangle, \langle \mu_2, c_2, \mu_2 \rangle$ be generative classes for the catenation of ba^{i_k} to w_{k-1} to yield $w_k = w_{k-1}ba^{i_k}$. If $\langle \mu_2 \rangle \geq_s \langle \mu_1, c_1, \mu_1 \rangle$, then the number of squares in the classes created by the catenation is at most $\min\{c_1, i_k\} + 1 \leq \min\{I(w_k)[\max -2], i_k\} + 1$. Moreover, if $\langle \mu_2 \rangle >_s \langle \mu_1, c_1, \mu_1 \rangle$, then $c_2 < c_1$.*

Proof. Let $w_{k-1} = a^{i_0}b\cdots ba^{i_{k-1}}$, and we have $\langle i_0,\ldots,i_{k-1}\rangle \geq_s \langle \ell_2,\mu_2,c_2,\mu_2\rangle$ and $\langle c_2,\mu_2\rangle \geq_s \langle \ell_1,\mu_1,c_1,\mu_1\rangle$ for some ℓ_1,ℓ_2. The catenation creates at most[1]

$$\min\{\ell_2,c_2\}+\min\{c_2,i_k\}-c_2+1 \tag{7}$$

squares in the class $\langle \mu_2,c_2,\mu_2\rangle$ due to (4).

We first consider the case when $\langle \mu_1,c_1,\mu_1\rangle$ is a *proper* suffix of $\langle \mu_2\rangle$. In counting the number of distinct squares to be created in the class $\langle \mu_1,c_1,\mu_1\rangle$, we should take into account the factor $\langle \ell_1,\mu_1,c_1,\mu_1,c_2\rangle$ of w_{k-1}. In order for the class to be generative, we have $c_2 < \min\{c_1,i_k\}$. Then at most $\min\{c_1,i_k\}-c_2$ squares in the class are created, and the subtraction term "$-c_2$" cancels (7). As a result, the catenation creates at most $\min\{c_1,i_k\}+1$ squares in these two classes. Moreover, this is upper bounded by $\min\{I(w_k)[\max-2],i_k\}+1$ since $I(w_k)$ contains two c_1's.

Next we consider the case of $\langle c_2,\mu_2\rangle = \langle \ell_1,\mu_1,c_1,\mu_1\rangle$ (see Fig. 2), that is, $c_2=\ell_1$ and $\langle \mu_2\rangle = \langle \mu_1,c_1,\mu_1\rangle$. It creates at most the following number of distinct squares in the class $\langle \mu_1,c_1,\mu_1\rangle$:

$$\begin{aligned}&\min\{c_2,c_1\}+\min\{c_1,i_k\}-c_1+1\\ &\quad-\max\big\{\min\{\ell_2,c_2,c_1\}+\min\{c_1,c_2,i_k\}-c_1+1,0\big\}.\end{aligned} \tag{8}$$

The subtraction term, due to (5), takes into account that $\langle \ell_2,\mu_1,c_1,\mu_1,c_2\rangle$ already appears in $\langle i_0,\ldots,i_{k-1}\rangle$. The number of distinct squares in these classes created by the catenation is given as the sum (7) + (8). The last subtraction term in (8) is 0 if and only if $\min\{\ell_2,c_2\}+\min\{c_2,i_k\}<c_1$. Then,

$$\begin{aligned}(7)+(8) &= (\min\{\ell_2,c_2\}+\min\{c_2,i_k\}+1-c_1)\\ &\quad+\min\{c_1,i_k\}+(\min\{c_2,c_1\}-c_1)+1 \leq \min\{c_1,i_k\}+1.\end{aligned}$$

If the term is positive, on the other hand, then we obtain

$$\begin{aligned}(7)+(8) &= (\min\{\ell_2,c_2\}+\min\{c_2,c_1\}-\min\{\ell_2,c_2,c_1\}-c_2)\\ &\quad+(\min\{c_2,i_k\}+\min\{c_1,i_k\}-\min\{c_2,c_1,i_k\})+1 \leq i_k+1.\end{aligned}$$

Now we prove that the sum is at most c_1+1, and it suffices to do so under the condition $c_1<i_k$. Then the sum is $(\min\{\ell_2,c_2\}+\min\{c_2,i_k\}-c_2-\min\{\ell_2,c_2,c_1\})+c_1+1$. If $\min\{\ell_2,c_2,c_1\}=\min\{\ell_2,c_2\}$, then the terms inside the parentheses amount to 0 and hence the sum is at most c_1+1. This condition must hold because if $\min\{\ell_2,c_2,c_1\}=c_1$, then the class $\langle \mu_1,c_1,\mu_1\rangle$ would have been already saturated in w_{k-1} so that it could not be generative. □

Now we develop the previous argument for arbitrary number of generative classes: $\langle \mu_m,c_m,\mu_m\rangle$, …, $\langle \mu_2,c_2,\mu_2\rangle$, $\langle \mu_1,c_1,\mu_1\rangle$ with $m\geq 3$ such that $\langle \mu_m,c_m,\mu_m\rangle >_s \cdots >_s \langle \mu_1,c_1,\mu_1\rangle$. Interestingly, *no matter how many generative classes are involved*, catenation creates at most $(\min\{I(w_k)[\max-3],i_k\}+1)+(\min\{I(w_k)[\max-2],i_k\}+1)$ squares. The next lemma enables us to divide the classes into two groups so that the classes in one group are responsible for the first term and those in the other are for the second term.

[1] Here we say "at most" because w_{k-1} may contain some squares in this class already.

Lemma 6. *Let* $\langle \mu_1, c_1, \mu_1 \rangle, \langle \mu_2, c_2, \mu_2 \rangle, \langle \mu_3, c_3, \mu_3 \rangle$ *be three generative classes of the catenation of* ba^{i_k} *to* w_{k-1} *to yield* $w_k = w_{k-1}ba^{i_k}$ *such that* $\langle \mu_3, c_3, \mu_3 \rangle >_s \langle \mu_2, c_2, \mu_2 \rangle >_s \langle \mu_1, c_1, \mu_1 \rangle$. *Then* $\langle \mu_3 \rangle >_s \langle \mu_1, c_1, \mu_1 \rangle$ *and* $c_3 < c_1$ *hold, and the number of squares in the classes* $\langle \mu_3, c_3, \mu_3 \rangle$ *and* $\langle \mu_1, c_1, \mu_1 \rangle$ *created by the catenation is at most* $\min\{c_1, i_k\} + 1 \leq \min\{I(w_k)[\max -2], i_k\} + 1$.

Proof. If $\langle \mu_3 \rangle >_s \langle \mu_1, c_1, \mu_1 \rangle$ did not hold, then, by Lemma 2, $\langle c_1, \mu_1, c_1, \mu_1, c_1 \rangle$ would be a factor of the coefficient sequence of w_{k-1}, that is, the class $\langle \mu_1, c_1, \mu_1 \rangle$ would be saturated in w_{k-1}, a contradiction. Thus, $\langle \mu_3 \rangle >_s \langle \mu_1, c_1, \mu_1 \rangle$ must hold. The other two results derive from this due to Lemma 5. □

Consider the catenation of ba^{i_k} to w_{k-1} from the right, and let $\langle \mu_{i_\ell}, c_{i_\ell}, \mu_{i_\ell} \rangle$, $\ldots$, $\langle \mu_{i_1}, c_{i_1}, \mu_{i_1} \rangle$ be its generative classes with $i_m > \cdots > i_1$. We say that they form a *(length-halving) chain* if for any $1 < j \leq \ell$, $\langle \mu_{i_j} \rangle \geq_s \langle \mu_{i_{j-1}}, c_{i_{j-1}}, \mu_{i_{j-1}} \rangle$. Lemmas 5 and 6 imply:

Lemma 7. *For any* $\ell \geq 1$, *if the catenation of* ba^{i_k} *to* w_{k-1} *involves* ℓ *generative classes* $\langle \mu_{i_\ell}, c_{i_\ell}, \mu_{i_\ell} \rangle$, $\ldots$, $\langle \mu_{i_1}, c_{i_1}, \mu_{i_1} \rangle$ *that satisfy* $\langle \mu_{i_\ell}, c_{i_\ell}, \mu_{i_\ell} \rangle >_s \cdots >_s \langle \mu_{i_1}, c_{i_1}, \mu_{i_1} \rangle$ *and also form a chain, then the catenation creates at most* $\min\{c_{i_1}, i_k\} + 1$ *squares in these classes.*

Lemma 6 enables us to divide the classes into (at most) two groups so as for the classes in each group to form a chain. The index-parity-based division: $\ldots, \langle \mu_4, c_4, \mu_4 \rangle, \langle \mu_2, c_2, \mu_2 \rangle$ and $\ldots, \langle \mu_3, c_3, \mu_3 \rangle, \langle \mu_1, c_1, \mu_1 \rangle$ is such a division. With Lemma 7, now we complete the proof that the catenation cannot create more than $\min\{I(w_k)[\max -3], i_k\} + \min\{I(w_k)[\max -2], i_k\} + 2$ squares.

4.3 Towards an Inductive Proof for General Words

Any word can be factorized into slopes. Given a word, a *proper* factor $\langle i_\ell, i_{\ell+1}, \ldots, i_{r-1}, i_r \rangle$ ($l > 0$ and $r < k$) of its coefficient sequence is called a (local) *minimum* (*maximum*) if $i_\ell > i_{\ell+1} = i_{\ell+2} = \cdots = i_{r-1} < i_r$ (resp. $i_\ell < i_{\ell+1} = \cdots = i_{r-1} > i_r$). Minima and maxima are collectively called *extrema*. It must be noted that by definition extrema are a *proper* factor so that the leftmost or rightmost coefficient of the given word cannot be a part of them. For $m \geq 0$, we say that a word is an m-extrema word if it contains *at most* m extrema. By $\mathcal{E}_m$, we denote the class of all m-extrema words.

We identify two minima $\langle i_{\ell_1}, i_{\ell_1+1}, \ldots, i_{r_1-1}, i_{r_1} \rangle, \langle i_{\ell_2}, i_{\ell_2+1}, \ldots, i_{r_2-1}, i_{r_2} \rangle$ if $r_1 - \ell_1 = r_2 - \ell_2$ and $i_{\ell_1+j} = i_{\ell_2+j}$ for any $0 \leq j \leq r_1 - \ell_1$; otherwise, we say they are *distinct*.

Although μ_i is a subsequence of integers, we consider it a word where each integer is a symbol. This notation is applied in the following lemma.

Lemma 8. *If a catenation involves two generative classes* $\langle \mu_1, c_1, \mu_1 \rangle, \langle \mu_2, c_2, \mu_2 \rangle$ *in different chains, and all minima of the resulting word are pairwise-distinct, then one of the following holds:*

1. $c_1 > c_2$, $\mu_2 = c_2^m c_1 c_2^j$, *and* $\mu_1 = c_2^j$ *for some* $j \geq 1$ *and* $m < j$;

2. $c_1 < c_2$, $\mu_2 = c_1^{2j+m+1}c_2c_1^{j+m}$, *and* $\mu_1 = c_1^jc_2c_1^{j+m}$ *for some* $j \geq 1$ *and* $m \geq 0$;
3. $c_1 \neq c_2$, $\mu_2 = d^jc_1c_2d^j$, *and* $\mu_1 = c_2d^j$ *for some* $j \geq 0$ *and coefficient* d *with* $d \leq c_1$ *and* $d < c_2$.

Theorem 3. *If all minima of a word* w_k *of length* n *with* k b*'s are pairwise-distinct, then* $\#\mathrm{Sq}(w_k) \leq \frac{2k-1}{2k+2}n$.

Proof. Let $w_k = a^{i_0}b\cdots ba^{i_k}$ and consider the catenation of ba^{i_k} to $w_{k-1} = a^{i_0}b\cdots ba^{i_{k-1}}$ to yield w_k. Assume two generative classes $\langle \mu_1, c_1, \mu_1\rangle, \langle \mu_2, c_2, \mu_2\rangle$ are involved in it, and moreover, they are in different chains. To them, Lemma 8 is applicable to represent these classes in three ways. Proofs for all these representations take the same strategy: spotting an coefficient i_j such that catenating ba^{i_j} creates so small number of squares that offsets the number of squares to be created by the catenation of ba^{i_k}. Therefore, in the following, we just examine the first representation.

We have that $\langle \ell, \mu_2, c_2, \mu_2\rangle$ is a suffix of the coefficient sequence $\langle i_0, \ldots, i_{k-1}\rangle$ of w_{k-1} and $\mu_2 = c_2^mc_1c_2^j$ for some coefficients ℓ, c_1, c_2 with $c_1 > c_2$ and $j \geq 1, m \geq 0$ with $j > m$. The right μ_2 is actually the sequence $\langle i_{k-j}, \ldots, i_{k-j+m-1}, i_{k-j+m}, \ldots, i_{k-1}\rangle$. Consider the successive catenations of $ba^{i_{k-j+m}}, \ldots, ba^{i_{k-1}}$ to $w_{k-j+m-1} = a^{i_0}b\cdots ba^{i_{k-j+m-1}}$. If $\ell \neq c_2$, then the first catenation creates $\max\{c_2 - \ell, 0\}$ squares in the class $\langle c_2^m, c_1, c_2^m\rangle$, which is its sole generative class. The catenation of i_k creates at most $\min\{c_2, i_k\} + \min\{\ell, c_2, i_k\} + 2$ squares. As a result, they introduce two additive terms. Moreover, due to $\ell \neq c_2$, each of other catenations involves just one chain. If $\ell = c_2$, then $\langle c_2^m, c_1, c_2^m\rangle$ is not generative any more, but instead the class $\langle c_2^jc_1c_2^m, c_2, c_2^jc_1c_2^m\rangle$ can be. If it is not, then no square is created, and this offsets one term brought by the catenation of ba^{i_k}. Otherwise, $\langle i_0, \ldots, i_{k-1}\rangle \geq_s \langle \ell', c_2^jc_1c_2^m, c_2, c_2^jc_1c_2^m\rangle$ for some $\ell' \geq 0$. The catenation of i_k creates at most $\min\{\ell', c_2\} + \min\{c_2, i_k\} - c_2 + 1$ squares in the class $\langle \mu_2, c_2, \mu_2\rangle$ and $c_2 + \min\{c_1, i_k\} - c_1 + 1 - (\min\{\ell', c_2\} + \min\{c_2, i_k\} - c_1 + 1)$, where the subtraction term is to avoid the double-counting (note $\langle \ell', \mu_1, c_1, \mu_1, c_2\rangle$ is in $\langle i_0, \ldots, i_{k-1}\rangle$). Thus, it creates at most $\min\{c_1, i_k\} + 1$ squares. □

As its corollary, we can verify the bound $\frac{2k-1}{2k+2}n$ for the word (1) by Fraenkel and Simpson, or more precisely, for its factors with k b's, since all of their minima are pairwise-distinct.

Corollary 2. *For any factor* $w_{\mathrm{fs},k}$ *of* w_{fs} *with* k b*'s,* $\#\mathrm{Sq}(w_{\mathrm{fs},k}) \leq \frac{2k-1}{2k+2}|w_{fs,k}|$.

Maxima-pairwise-distinct variants of Lemma 8 and Theorem 3 hold. As for the variant of the lemma, all inequalities must be inverted. From them, the next result holds.

Corollary 3. *For any word* $w_k \in \mathcal{E}_3$ *with* k b*'s,* $\#\mathrm{Sq}(w_k) \leq \frac{2k-1}{2k+2}|w_k|$.

Corollary 4. *For any* $k \leq 6$ *and word* w_k *with* k b*'s,* $\#\mathrm{Sq}(w_k) \leq \frac{2k-1}{2k+2}|w_k|$.

Proof. It suffices to observe that, for any $k \geq 3$, words with k b's can contain at most $k - 3$ extrema. Then this immediately follows from Corollary 3. □

The more classes are involved, the more strictly the structure of w_{k-1}, to which we catenate ba^{i_k}, is restricted. In fact, we can easily show that for $k \leq 9$, either the catenation involves just one chain or all minima (or maxima) of the resulting word are pairwise-distinct.

Proposition 2. *For any $k \leq 9$ and word w_k with k b's, $\#\mathrm{Sq}(w_k) \leq \frac{2k-1}{2k+2}|w_k|$.*

5 Conclusions

Our results are partial steps in showing Conjecture 1. However, we identified several ways to approach this conjecture. For instance, one may follow the technique in which we examine a word as a sequence of slopes, and try to identify how the number of squares increases when the words have non pairwise distinct minima (maxima). Nevertheless, it may be the case that a direct inductive proof with respect to the number of b's would validate the conjecture; using the generative classes with respect to catenation we only analyzed the cases when this number is at most 9, but it is our hope that our method can be generalized.

Finally, we discussed only the case of binary words. It seems unlikely that the tools we developed could be used directly to obtain upper bounds on the number of squares in words over larger alphabets.

Acknowledgement. We gratefully acknowledge helpful discussions with Florence Linez, Robert Mercaş, and Mike Müller. Mike Müller kindly implemented a computer program for the experimental verification of Conjecture 1. The research was partially supported by the NSF grants No. DMS-0900671 and CCF-1117254 and the NIH grant R01 GM109459-01 to N. J., by the DFG grant 596676 to F. M., and by the HIIT Pump Priming Project Grant 902184/T30606 and the Academy of Finland, Postdoctoral Researcher Grant 13266670/T30606 to S. S.

References

1. Fraenkel, A.S., Simpson, J.: How many squares can a string contain? Journal of Combinatorial Theory, Series A 82, 112–120 (1998)
2. Deza, A., Franek, F., Jiang, M.: A d-step approach for distinct squares in strings. In: Giancarlo, R., Manzini, G. (eds.) CPM 2011. LNCS, vol. 6661, pp. 77–89. Springer, Heidelberg (2011)
3. Ilie, L.: A note on the number of squares in a word. Theoretical Computer Science 380, 373–376 (2007)
4. Fan, K., Puglisi, S.J., Smyth, W.F., Turpin, A.: A new periodicity lemma. SIAM Journal of Discrete Mathematics 20(3), 656–668 (2005)
5. 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)
6. Kopylova, E., Smyth, W.F.: The three squares lemma revisited. Journal of Discrete Algorithms 11, 3–14 (2012)