Squares in Partial Words∗

F. Blanchet-Sadri1 Yang Jiao2 John M. Machacek3
J. D. Quigley4 Xufan Zhang5

March 4, 2014

Abstract

We investigate the number of positions that do not start a square, the number of square occurrences, and the number of distinct squares in partial words, i.e., sequences that may have undefined positions called holes. We show that the limit of the ratio of the maximum number of positions not starting a square in a binary partial word with $h$ holes over its length $n$ is $15/31$ and the limit of the ratio of the minimum number of square occurrences in a binary partial word with $h$ holes over its length $n$ is $103/187$, provided the limit of $h/n$ is 0. Both limits turn out to match with the known limits for binary full words (those without holes). We prove another surprising result that the maximal proportion of defined positions that are square-free to the number of defined positions in a binary partial word with $h$ holes of length $n$ is $1/2$, provided the limit of $h/n$ is in the interval $[1/11, 1)$. We also give a $2k^h$ tight bound on the number of rightmost occurrences of squares per position in a $k$-ary partial word with $h$ holes. In addition, we provide a more detailed analysis than earlier ones for the maximum number of distinct squares in a one-hole partial word of length $n$ over an alphabet of size $k$, bound that is independent of $k$.

∗This material is based upon work supported by the National Science Foundation under Grant No. DMS–1060775. Part of this paper was presented at DLT 2012 [3].

1Department of Computer Science, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402–6170, USA, blanchet@uncg.edu
2Department of Mathematical Sciences, Carnegie Mellon University, Wean Hall 6113, Pittsburgh, PA 15213, USA
3Department of Mathematics, Michigan State University, 619 Red Cedar Road, East Lansing, MI 48824, USA
4Department of Mathematics, University of Illinois at Urbana-Champaign, 1409 W. Green Street, Urbana, IL 61801, USA
5Department of Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544–1000, USA
Keywords: Combinatorics on words; Partial words; Square positions; Distinct squares; Square occurrences.

1 Introduction

A square in a word consists of two adjacent occurrences of a subword. We refer to the following example of a binary word to illustrate the concepts we will be talking about:

\[
\begin{array}{cccccccccccccc}
    & i & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 \\
    w_1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
    sd(w_1, i) & 0 & 1 & 1 & 1 & 1 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\end{array}
\]

Here, 010010 = (010)\(^2\) is an instance of a square that occurs twice in the word \(w_1\) of length 19. We have \(w_1[3..8] = w_1[10..15] = 010010\), which yields two square occurrences positioned at 3 and 10 (we also say that positions 3 and 10 are square positions since they start squares). The sequence \(sd(w_1, 0) \cdots sd(w_1, 18)\) is such that each \(sd(w_1, i)\) represents the number of distinct squares whose rightmost occurrence begins at position \(i\) in \(w_1\) (there are 11 distinct squares in \(w_1\)).

A question that has been investigated is “How many distinct squares are there in a \(k\)-ary word of length \(n\)?” Here each square is counted only once. The answer is \(O(n)\), and Fraenkel and Simpson [7] showed in 1998 that this number is at most \(2n\) since at most two distinct squares can have their rightmost occurrence starting at the same position. The bound \(2n\) has a simpler proof by Ilie in 2005 [10] and improved to \(2n - \Theta(\log n)\) in 2007 [11]. A conjecture, supported by computations, states that the number of distinct squares in a \(k\)-ary word of length \(n\) is at most \(n\). The upper bound \(n\) is optimal since there is a construction with asymptotically \(n\) distinct squares of which the example word \(w_1\) is the second iteration [7], and the conjecture is believed to be very difficult to prove. Upper bounds on the maximum number of consecutive 2’s in \(sd_k(w, 0) \cdots sd_k(w, n - 1)\) were used to improve the bound of \(2n\) to \(2n - \Theta(\log n)\).

Another question that has been investigated is “How many square occurrences are there in a word of length \(n\)?”. In [12], Kucherov, Ochem, and Rao studied the number of square occurrences in a binary word. More specifically, they studied a question that was left open in [6, 7]: “What is the minimal limit proportion of square occurrences in an infinite binary word?”.

Kucherov et al. showed that this number is, in the limit, a constant fraction of the word length, and gave a very good estimation of this constant. Later on, in [15], Ochem and Rao proved that the limit of the ratio of the minimal number of square occurrences in a binary word over its length is \(\frac{103}{187} \approx 0.55080\). Furthermore in [9], Harju, Kärki, and Nowotka considered
the number \(\text{sp}(w)\) of positions that do not start a square in a binary word \(w\). Letting \(\text{sp}(n)\) denote the maximum of the \(\text{sp}(w)\)'s, where \(|w| = n\), they showed that \(\lim \text{sp}(n)/n = \frac{15}{31}\).

Blanchet-Sadri, Merca¸s, and Scott [5] started investigating the problem of counting distinct squares in partial words of length \(n\) over a \(k\)-letter alphabet, a problem that revealed surprising results. Partial words are sequences that may have undefined positions, called holes and denoted by \(\diamond\)'s, that match, or are compatible with, any letter of the alphabet. In this context, a square has the form \(uv\) with \(u\) and \(v\) compatible, and consequently, such square is compatible with a number of full words (those without holes) over the alphabet that are squares. For example, \(w = 01\diamond01\diamond01\) contains 9 squares over \(\{0,1\}\) since \(0^2, 1^2, (10)^2, (001)^2, (010)^2, (101)^2, (110)^2\), and \((110)^2\) are the 9 full words that are squares compatible with factors of \(w\). It was shown that, unlike for full words, the number of distinct squares in partial words can grow polynomially with respect to \(k\), and bounds, dependent on \(k\), were given in a few cases.

The one-hole case behaves very differently from the zero-hole case. It was proved in [5] that for partial words with one hole, there may be more than two squares that have their rightmost occurrence at the same position, and that if such is the case, then the hole is in the shortest square. This is the case with \(0\diamond0101\diamond01\) that has three squares with rightmost occurrences at position 0, i.e., \((00)^2\), \((01)^2\), and \((01001)^2\). In [8], Halava, Harju, and Kärki showed that the maximum number of the rightmost occurrences of squares per position in a partial word with one hole is \(2k\).

In this paper, we prove results on the number of positions that do not start a square, the number of square occurrences, and the number of distinct squares in a partial word of a given length. For clarity, each of Sections 3, 4 and 5 has been divided into two subsections titled “Binary alphabet” and “Arbitrary alphabet”. When dealing with an arbitrary alphabet, we refer to its size as \(k\).

The contents of our paper are as follows: In Section 2, we review some basic preliminaries on partial words. In Section 3, letting \(\text{sp}_h(n)\) be the maximum number of positions not starting a square for binary partial words with \(h\) holes of length \(n\), we show that \(\lim \text{sp}_h(n)/n = \frac{15}{31}\) provided the limit of \(h/n\) is 0. We prove another surprising result that the maximal proportion of defined positions that are square-free to the number of defined positions in a binary partial word with \(h\) holes of length \(n\) is equal to \(\frac{1}{2}\), provided the limit of \(h/n\) is in the interval \([\frac{1}{11}, 1]\). In Section 4, we extend Halava et al.’s \(2k\) tight bound on the number of rightmost occurrences of squares per position in the case of a \(k\)-ary partial word with one hole to a \(2k^h\) tight bound.
in the case of a \( k \)-ary partial word with \( h \) holes. We also give a more detailed analysis than the ones in [4, 8] for an upper bound on the number of distinct squares in a one-hole partial word of length \( n \) over a \( k \)-letter alphabet, bound that is independent of the alphabet size \( k \). In Section 5, letting \( \mathsf{so}_h(n) \) be the minimum number of square occurrences in a binary partial word of length \( n \) with \( h \) holes, we show that \( \lim \mathsf{so}_h(n)/n = \frac{103}{187} \approx 0.55080 \) provided the limit of \( h/n \) is zero. To do this, we modify Kucherov et al.'s construction for the upper bound, which is based on a certain pattern of length 187 that they discovered while computing long words that achieve the minimum number of square occurrences for their length. We exhibit a constant \( C \) such that if the limit of \( h/n \) is bounded by \( C \), then \( \lim \mathsf{so}_h(n)/n \approx 0.55081 \). We also prove a result about the minimal limit proportion of square occurrences in a \( k \)-ary partial word. Finally in Section 6, we conclude with some conjectures and open problems.

2 Preliminaries

For more background material on partial words, we refer the reader to [2].

We let \( A_\circ \) denote an alphabet \( A \) along with the hole symbol \( \circ \), i.e., \( A_\circ = A \cup \{\circ\} \). The set of all finite sequences constructed from elements of \( A \) (resp., \( A_\circ \)) is denoted by \( A^* \) (resp., \( A_\circ^* \)) and the set of all finite sequences of length \( n \) constructed from elements of \( A \) (resp., \( A_\circ \)) by \( A^n \) (resp., \( A_\circ^n \)).

A full word over \( A \) is any \( v \in A^* \) and a partial word over \( A \) is any \( v \in A_\circ^* \) (note that every full word is also a partial word that does not have any \( \circ \)). We denote the empty word consisting of no symbols by \( \varepsilon \).

We use \( |v| \) to denote the length of partial word \( v \) or the number of symbols in \( v \). The set of positions of \( v \) which are holes is denoted by \( H(v) \) and the set of the remaining positions by \( D(v) \).

We say that \( u \) is a factor of a partial word \( v \) if there exist \( x, y \) such that \( v = xuy \). The factor of \( v \) starting at position \( i \) and ending at position \( j \) will be denoted by \( v[i..j] \), while the factor of \( v \) starting at position \( i \) and ending at position \( j - 1 \) will be denoted by either \( v[i..j - 1] \) or \( v[i..j] \).

Two partial words \( u \) and \( v \) of same length are compatible, denoted by \( u \uparrow v \), if they are equal at all positions in \( D(u) \cap D(v) \); \( u \) is contained in \( v \), denoted by \( u \subset v \), if \( u \) is equal to \( v \) at all positions in \( D(u) \). The least upper bound \( u \lor v \) of two compatible partial words \( u \) and \( v \) satisfies \( u \subset (u \lor v) \), \( v \subset (u \lor v) \), and \( D(u \lor v) = D(u) \cup D(v) \). A square in a partial word is a factor of the form \( uv \), where \( u \) and \( v \) are compatible. We call \( u \lor v \) the root of the square.
Given \( v = a_0 \cdots a_{n-1} \), where \( a_i \in A_0 \), the reverse of \( v \) is \( \text{rev}(v) = a_{n-1} \cdots a_0 \). In case \( A = \{0, 1\} \), the complement of \( v \) is \( \overline{v} = \overline{a_0} \cdots \overline{a_{n-1}} \), where \( \overline{0} = 1 \), \( \overline{1} = 0 \), and \( \overline{\circ} = \circ \). In this paper, we denote the binary alphabet \( \{0, 1\} \) by \( B \).

3 Square Positions

Given \( w = a_0 a_1 \cdots a_{n-1} \), where each \( a_i \in A \circ \), position \( i \) starts a square if \( a_i a_{i+1} \cdots a_{i+j-1} \uparrow a_{i+j} a_{i+j+1} \cdots a_{i+2j-1} \) for some \( j \). If position \( i \) does not start a square, then \( i \) is called square-free. Given an occurrence of a factor \( u \) of \( w \), let \( \text{sp}_w(u) \) be the number of positions in \( u \) that are square-free in \( w \) (when referring to \( \text{sp}_w(u) \), the occurrence of \( u \) in \( w \) is implicitly assumed without any risk of confusion).

Consistent with [9], \( w \) is strong if \( \text{sp}_w(u) \geq \frac{|u|}{2} \) for every nonempty prefix \( u \) of \( w \). In the case \( u = w \), we let \( \text{sp}(w) = \text{sp}_w(w) \). A simple observation of a necessary condition for a partial word to be strong is that every nonempty prefix of a strong partial word must be strong. Also, the min-factor \( m(w) \) of \( w \) is the shortest prefix \( u \) of \( w \) such that \( \text{sp}_w(u) < \frac{|u|}{2} \), if it exists, and the min-decomposition of \( w \) is the factorization \( w = w_1 w_2 \cdots w_r w_{r+1} \) where \( w_i = m(w_i \cdots w_{r+1}) \) for \( 1 \leq i \leq r \) and \( w_{r+1} \) does not have a min-factor.

We also let

\[
\text{sp}_{h,k}(n) = \max \{\text{sp}(w) : w \in A_0^*, |w| = n, \|H(w)\| = h\},
\]

where \( A \) is any alphabet of size \( k \). To simplify the notation, we will abbreviate \( \text{sp}_{h,2}(n) \) by \( \text{sp}_{h}(n) \).

3.1 Binary alphabet

We first characterize all strong binary partial words and look at the asymptotic behavior of the ratio \( \frac{\text{sp}_{h}(n)}{n} \).

Lemma 1. Given a binary partial word \( w = u \circ v \in B_0^* \), with \( u \) a full word and \( v \neq \epsilon \), \( \text{sp}_w(u \circ) < \frac{|u|+1}{2} \).

Proof. We proceed by induction on \(|u|\). For base cases \( 0 \leq |u| \leq 4 \), the result is clear. For instance, if \(|u| = 4\), then the last position of \( u \) and the position of the hole start squares. Since \(|u| = 4\), there must be another square in one of the first three positions in \( u \). Suppose that for some \( n \geq 4 \) the result holds for all \( u \) with \(|u| \leq n\). We show it holds for \( t \) with \(|t| = n + 1\), and we write \( t \) as \( 0u \) or \( 1u \) for some \( u \) with \(|u| = n\).
First consider the case where $sp_w(u\diamond) < \frac{|u|}{2}$. Then, at worst, the first position of $t$ can be square-free. Nonetheless, we still have $sp_w(t\diamond) < \frac{|u|+1}{2}$. The only other case to consider is when $sp_w(u\diamond) = \frac{|u|}{2}$. Either $u$ is strong or it is not. If $u$ is strong, since computer results yield no such $u$ of length less than 8, $|u| \geq 8$ and hence [9] proves that such a $u$ either starts with 010 or 101; in either case, we add a square to the first position of $t$. If $u$ is not strong, the min-decomposition of $u$ is nontrivial and is given by $u = u_1 \cdots u_{r+1}$, where $u_{r+1}$ is strong. By definition, $sp_w(u_i) < \frac{|u|}{2}$ for $i \leq r$. By the inductive hypothesis, $sp_w(u_{r+1}\diamond) < \frac{|u_{r+1}|+1}{2}$. We thus obtain a contradiction with $sp_w(u\diamond) = \frac{|u|}{2}$.

In [9], it was shown that there are 382 strong binary full words with the longest having length 37.

**Theorem 1.** There are 95 strong binary partial words with one hole the longest of which has length 37, and each is of the form $w = u\diamond$ for a strong full word $u$. Furthermore, there are no strong binary partial words with more than one hole.

**Proof.** From Lemma 1, we can see for a binary partial word to be strong and have a hole that the hole must be the last character. This also rules out the possibility of any strong binary partial word with more than one hole. Thus any such partial word must be of the form $w = u\diamond$ for a full word $u$, and certainly $u$ must be a strong full word. A computer check can verify that there are only 95 such partial words and that the longest has length 37.

As long as the number of holes in a partial word grows asymptotically slower than its length, we can realize the same limit as for full words.

**Theorem 2.** Let $\{h_n\}$ be an integer sequence such that $h_n \leq n$ for all $n$. If
\[
\lim_{n \to \infty} \frac{h_n}{n} = 0, \text{ then } \lim_{n \to \infty} \frac{sp_{h_n}(n)}{n} = \frac{15}{31}.
\]

**Proof.** From [9], the limit is $\frac{15}{31}$ in the zero-hole case, and clearly $sp_h(n) \leq sp_0(n)$ for all $0 \leq h \leq n$. Also if we consider $w = \diamond^hu$ for a full word $u$ where $|u| = n - h$ and $sp(u) = sp_0(n - h)$, we can see that $sp_h(n) \geq sp_0(n - h)$. So
\[
\lim_{n \to \infty} \frac{sp_0(n - h_n)}{n} \leq \lim_{n \to \infty} \frac{sp_{h_n}(n)}{n} \leq \lim_{n \to \infty} \frac{sp_0(n)}{n}.
\]
Noting that $\frac{sp_0(n - h_n)}{n} = \left(\frac{n - h_n}{n}\right) \left(\frac{sp_0(n - h_n)}{n - h_n}\right)$, we can see our result holds. \qed
We now discuss a slightly different problem. For any partial word $w$, we define $\delta(w)$ to be the size of the set \( \{ i : i \in D(w), \text{position } i \text{ is square-free} \} \).

Then we define

\[
\text{rsp}_{h}(n) = \max \left\{ \frac{\delta(w)}{\|D(w)\|} : w \in B^*_\circ, |w| = n, \|H(w)\| = h \right\}.
\]

Theorem 2 implies that if \( \{h_n\} \) is an integer sequence such that $h_n < n$ for all $n$, and if \( \lim_{n \to \infty} \frac{h_n}{n} = 0 \), then

\[
\lim_{n \to \infty} \text{rsp}_{h_n}(n) = \frac{15}{31}.
\]

This is because when \( \lim_{n \to \infty} \frac{h_n}{n} = 0 \), the ratios of $\circ$s in the word and in the square-free positions go to 0, and \( \frac{\text{sp}_{h_n}(n)}{n}, \text{rsp}_{h_n}(n) \) tend to the same limit.

Theorems 3 and 4 discuss the case when \( \lim_{n \to \infty} \frac{h_n}{n} > 0 \).

**Theorem 3.** Let \( \{h_n\} \) be an integer sequence such that $h_n < n$ for all $n$. If \( \lim_{n \to \infty} \frac{h_n}{n} = \ell \) for some $\ell \in [0,1)$, then

\[
\limsup_{n \to \infty} \text{rsp}_{h_n}(n) \leq \frac{1}{2}.
\]

**Proof.** We scan any partial word $u$ of length $n$ from left to right. Denote the current position as $j$.

If $j \in H(u)$, it does not contribute to $\delta(u)$ or $D(u)$, so we can skip it. Assume $j \in D(u)$. Let $l$ be the maximum number such that $u[j..l]$ does not contain a $\circ$. We denote $u[j..l]$ as $w$, and we take its min-decomposition

\[
w = w_1 w_2 \cdots w_r w_{r+1}
\]

where for any integer $i \in [1, r]$, the positions in $w_i$ that are square-free in $u$ are fewer than $|w_i|/2$, and $w_{r+1}$ is strong. So we only need to consider the strong full word $w_{r+1}$. There are only 382 strong full words, each of which has length at most 37.

Suppose $n - l \geq 2$, otherwise $w_{r+1}$ is at the end of $u$ and does not affect the limit as $n \to \infty$, because $\ell < 1$ implies $n - h_n \to \infty$ when $n \to \infty$. By our choice of $l$, $u[l+1] = \circ$. By Lemma 1, given any strong full word $v$, for any
word \( v' \) among \( v\circ\circ, v\circ1, v\circ0, v\circ1 \), the number of positions in \( v \) that are square-free in \( v' \) is less than or equal to \(|v|/2 \). Since these are the only possibilities, we know that the number of positions in \( w_{r+1} \) that are square-free in \( u \) is less than or equal to \(|w_{r+1}|/2 \). \( \square \)

**Theorem 4.** Let \( \{h_n\} \) be an integer sequence such that \( h_n < n \) and \( h_n/n \geq \frac{1}{11} \) for all sufficiently large \( n \). If \( \lim_{n \to \infty} \frac{h_n}{n} = \ell \) for some \( \ell \in \left[ \frac{1}{11}, 1 \right) \), then

\[
\liminf_{n \to \infty} \text{rsp}_{h_n}(n) \geq \frac{1}{2}.
\]

**Proof.** Define the alphabet \( A = \{1, 2, 3\} \). By [16], there exists a sequence \( (u_r)_{r \geq 0} \) of square-free words over \( A \) such that \( u_r \) is a prefix of \( u_{r+1} \), and 313 and 212 do not appear in any \( u_r \).

Define \( w_1, w_2, w_3 \in B^*_5 \) as

\[
\begin{align*}
    w_1 &= 010011101100010011\circ10110011101100, \\
    w_2 &= 010011101100010011\circ101100\circ01001100010011\circ10110011101100, \\
    w_3 &= 01001100010011\circ101100.
\end{align*}
\]

Notice that \(|w_1| = 34, |w_3| = 22 \) and \(|w_2| = 22 + 34 = 56 \). Also note that no \( w_i \) is compatible with a prefix of \( w_j \) for \( i \neq j \).

We define a morphism \( \phi : A^* \to B^*_5 \) by mapping each \( i \in A \) to \( w_i \). Let \( v_r = \phi(u_r) \). We claim that if \( v_r[j..j+1] = 01 \) or 10, then \( j \) is square-free in \( v_r \).

Assume the claim for now. For all \( j \in D(v_r) \) such that \( v_r[j..j+1] = 00, 11, 0\circ \) or \( 1\circ \), \( j \) starts a square in \( v_r \). We check that exactly half of \( D(w_i) \) are such positions, and \( v_r \) is a concatenation of \( w_i \)'s, so \( \delta(v_r) = \frac{1}{2} \).

By examining the density of \( \circ \)'s in \( w_1, w_2, w_3 \), we see that \( \|H(v_r)\|/|v_r| \leq \frac{1}{11} \). Since we assume \( h_n/n \geq \frac{1}{11} \) for all sufficiently large \( n \), we can choose an appropriate \( v_r \), and, if necessary, add some \( \circ \)'s and possibly some finite number of letters in front of \( v_r \), to obtain a witness of \( \text{rsp}_{h_n}(n) \). Since \( n - h_n \to \infty \) as \( n \to \infty \), the added letters do not contribute to the limit, and hence \( \liminf_{n \to \infty} \text{rsp}_{h_n}(n) \geq \frac{1}{2} \).

We now prove our claim. We define some notions that will be useful. Suppose \( u \in A^* \), and for some position \( j \) in \( \phi(u) \), define its *preimage* \( \text{pre}(j) \) to be the unique letter \( l \in A \) in \( u \) whose image in \( \phi(u) \) includes position \( j \).
Also define the *internal position* of \( j \) to be the unique relative position of \( j \) inside \( \text{w}_{\text{pre}(j)} \). For example, let \( u = 13 \), then \( \phi(u) = w_1w_3 = 010011101100010011101100111010011101100010011 \).

Consider the last position \( j = 55 \) in \( \phi(u) \). Then \( \text{pre}(j) = 3 \) since it is in the image of letter 3, and the internal position of \( j \) is 21, which is the last position in \( w_3 \).

We can verify the claim for all squares with root-length \( \leq 2|w_2| + 2 \) by computer. Otherwise, suppose we have a square \( pq \) of root-length at least \( 2|w_2| + 4 \) in \( v_r \), where \( p \uparrow q \) and \( p \) starts with 01 or 10. Each of \( p \) and \( q \) must contain a factor of the form \( w_c01 \) for some \( c \). Such words can be easily verified by computer to be incompatible with any factor of \( v_r \), except for positions \( j \) with \( \text{pre}(j) = c \) and internal position 0, i.e., \( v_r[j..j+|w_c|+2) = w_c01 \).

Now it is easy to see that the square \( pq \) is a factor of \( w_cxw_dxw_e \), for some \( c,d,e \), with a centre within \( w_d \) and with \( x \) being a concatenation of some \( w_i \)’s. Since \( u_r \) is square-free, we have \( c \neq d \) and \( d \neq e \). We only need to find decompositions of \( w_d = yz \) so that \( y \) is a prefix of \( w_e \) and \( z \) is a suffix of \( w_c \). It is easy to check that they exist only for \( d = 1 \) and \( c = e = 2 \), i.e., \( 2x'1x'2 \) is a factor in \( u_r \) for some \( x' \in A^* \). With 212, 313, 22 and 11 being forbidden in \( u_r \), it is a clear contradiction.

**Corollary 1.** Under the assumptions of Theorem 4,

\[
\lim_{n \to \infty} \frac{\text{sp}_{h_n}(n)}{n} = \frac{1}{2}.
\]

### 3.2 Arbitrary alphabet

We prove a result about the asymptotic behavior of the ratio \( \frac{\text{sp}_{h_n,k}(n)}{n} \) in the \( k \)-ary case, where \( k \geq 3 \). By [13], there exists a \( k \)-ary square-free full word for all \( k \geq 3 \). Therefore, \( \text{sp}_{0,k}(n) = n \) for all \( k \geq 3 \).

**Proposition 1.** Let \( \{h_n\} \) be an integer sequence such that \( h_n \leq n \). For \( k \geq 3 \),

\[
\lim_{n \to \infty} \frac{\text{sp}_{h_n,k}(n)}{n} = \lim_{n \to \infty} \frac{n - h_n}{n}.
\]

**Proof.** To see that \( \lim_{n \to \infty} \frac{n - h_n}{n} \) is a lower bound, for each \( n \), we can construct a partial word \( w = \diamond^{h_n} w' \) of length \( n \), where \( w' \) is a \( k \)-ary square-free full word. Then each \( \diamond \) starts a square, and since all of the \( \diamond \)'s occur before \( w' \),
every position in \( w' \) is square-free. Thus, \( sp(w) = n - h_n \) and \( sp_{h_n,k}(n) \geq n - h_n \).

To see that \( \lim_{n \to \infty} \frac{n - h_n}{n} \) is an upper bound, we show that \( sp(w') \leq sp(w) \) for any \( w' \) with the same length and same number of holes as \( w \). Note that the only positions \( i \) which start squares in \( w \) are positions such that \( w[i] = \diamond \). In \( w' \), every \( \diamond \) which does not occur at the last position must start a square. If a \( \diamond \) occurs at the last position and \( h_n < n \), there must be some letter followed by a \( \diamond \), so the position of the letter starts a square. If a \( \diamond \) occurs at the last position and \( h_n = n \), then \( w = \diamond^h_0 = w'' \). Thus, \( sp(w'') \leq sp(w) \) and \( sp_{h_n,k}(n) \leq n - h_n. \)

\[ \square \]

4 Distinct Squares

We consider the number \( sd_k(w) \) of distinct squares in a \( k \)-ary partial word \( w \). In doing this, we count the number of distinct full squares compatible with factors of \( w \), i.e., for each factor \( yz \) with \( y \uparrow z \), we count each full word \( x^2 \) such that \( yz \subset x^2 \). In this section only, we say that a square full word \( x^2 \) is at position \( i \) in a partial word \( w \) if the rightmost factor in \( w \) that \( x^2 \) is compatible with starts at position \( i \) in \( w \).

Let \( sd_k(w, i) \) be the number of distinct squares in \( w \) at position \( i \) for \( 0 \leq i < |w| \). For example if \( w = 01\diamond 10 \), then \( sd_2(w, 0) = 1 \) since \( (01\diamond 1) \subset (01)^2 \), \( sd_2(w, 1) = 1 \) since \( (1\diamond 10) \subset (10)^2 \), \( sd_2(w, 2) = 1 \) since \( (\diamond 1) \subset 1^2 \), \( sd_2(w, 3) = 0 \), and \( sd_2(w, 4) = 0 \). Let

\[ \Delta sd_k(w) = \max\{sd_k(w) - sd_k(\hat{w}) : w \subset \hat{w}, \hat{w} \in A_0^*, |H(\hat{w})| = |H(w)| - 1 \}. \]

Here \( \hat{w} \) is a strengthening of \( w \), which is a partial word that comes from filling in any hole in \( w \) with a letter from our \( k \)-sized alphabet \( A \). For example, let \( w = 0010\diamond 01 \), \( \hat{w}_0 = 0010001 \), and \( \hat{w}_1 = 0010101 \). Then \( w \) has 3 distinct squares \( 0^2, (01)^2, (10)^2 \); \( \hat{w}_0 \) has 1 square \( 0^2 \); \( \hat{w}_1 \) has the same 3 squares as \( w \). So \( \Delta sd_2(w) = 2 \).

We let

\[ sd_{h,k}(n) = \max\{sd_k(w) : w \in A_0^*, |w| = n, |H(w)| = h \}, \]

where \( A \) is any alphabet of size \( k \). We also let

\[ \Delta_{h,k}(n) = \max\{\Delta sd_k(w) : w \in A_0^*, |w| = n, |H(w)| = h \}, \]

\[ \Gamma_{h,k}(n) = \max\{sd_k(w) - sd_k(\hat{w}) : w \subset \hat{w}, \hat{w} \in A^n, |H(w)| = h \}, \]
where $A$ is any alphabet of size $k$. Here $\hat{w}$ is a completion of $w$, which is a full word that comes from filling in all holes of $w$ with letters from our alphabet. To simplify the notation, we will abbreviate
\[
\text{sd}_2(w), \text{sd}_2(w, i), \Delta \text{sd}_2(w), \text{sd}_{h, 2}(n), \Delta_{h, 2}(n), \Gamma_{h, 2}(n),
\]
by dropping the subscript 2.

The following theorem follows immediately from the definition of $\Gamma_{h, k}(n)$.

**Theorem 5.** The number of distinct squares in a $k$-ary partial word of length $n$ with $h$ holes is bounded above by $\text{sd}_0, k(n) + \Gamma_{h, k}(n)$.

### 4.1 Binary alphabet

First, we give some lower bounds on $\Gamma_1(n)$ and $\Delta_1(n)$.

**Proposition 2.** For $n \geq 8$, $\Gamma_1(n) = \Delta_1(n) \geq \left\lfloor \frac{n-1}{2} \right\rfloor$.

**Proof.** We construct a class of binary partial words $\{u_0, u_1, u_2, u_3\}$, each of which has exactly one hole, and $|u_r| = 4q + r$ for $0 \leq r < 4$ and $q \geq 2$. We enumerate the distinct squares in each $u_r$, and for each square we account for the effect of the hole in $u_r$. This allows us to compute both $\text{sd}(u_r)$ and $\Delta_{\text{sd}}(u_r)$.

For $u_0 = 0^{q-1}10^{q-2}010^q10^{q-1}$, we have the following squares:

- $\left\lfloor \frac{q}{2} \right\rfloor$ squares of the form $(0^s)^2$ for $1 \leq s \leq \left\lfloor \frac{q}{2} \right\rfloor$, none of which are dependent on the hole;
- $q$ squares of the form $(0^s10^t)^2$ for $s + t = q - 1$ with $s, t \geq 0$, each of which requires the hole acting as a 0;
- $q - 1$ squares of the form $(0^s10^t)^2$ for $s + t = q$ with $s, t \geq 1$, each of which requires the hole acting as a 0;
- 2 squares $1^2$ and $(0^{q-2}1)^2$, each of which requires the hole acting as a 1. (Note that these two squares coincide for $q = 2$; however, this does not influence the final result.)

For the partial words $u_1 = 0^{q-1}10^{q-1}00^q10^q$, $u_2 = 0^{q-1}10^{q-1}00^q10^{q+1}$, and $u_3 = 010^q10^{q-2}010^q10^{q-1}$, we can similarly argue on the occurrences of squares. We obtain that $\Delta_{\text{sd}}(u_r) = \left\lfloor \frac{|u_r| - 1}{2} \right\rfloor$ for $0 \leq r < 4$. \qed
Next, using the lower bound construction from [7], let
\[ q(i) = 0^{i+1}10^i10^{i+1} \]
for \( i \geq 1 \). Letting \( Q = q(1) \cdots q(m) \), \( Q \) has length \( \frac{3m^2 + 13m}{2} \) and \( \text{sd}(Q) = \frac{3m^2 + 7m}{2} + \left\lceil \frac{m+1}{2} \right\rceil - 3 \).

**Proposition 3.** For \( n = \frac{3m^2 + 13m}{2} \) with \( m \geq 2 \), there exists a partial word with one hole of length \( n \) with \( 3m^2 + 9m \) distinct squares.

**Proof.** We modify the construction from [7] for partial words with one hole. Define \( Q_0 = q(1) \cdots q(m-2)0m10^{m-1}q \). We count the additional squares in \( Q_0 \) that do not occur in \( Q \):

\[ \left\lceil \frac{m-1}{2} \right\rceil \text{ squares of the form } (0^p10) \text{ with } p = \left\lceil \frac{m+3}{2} \right\rceil, \ldots, m \] and 3 squares \((0^m1)^2\), \((0^m10)^2\), \((0^m-10^m1)^2\). So \( \text{sd}(Q_0) = \text{sd}(Q) + \left\lceil \frac{m-1}{2} \right\rceil + 3 = \frac{3m^2 + 9m}{2} \).

**Proposition 4.** The following lower bounds hold:

1. For \( 1 \leq h < 2\left\lfloor \frac{n}{4} \right\rfloor - 1 \), we have \( \Delta_h(n) \geq 2\left\lceil \frac{n}{4} \right\rceil - 1 \).
2. For \( 1 < h \leq n - 2 \), we have \( \Delta_h(n) \geq 2\left\lceil \frac{h}{2} \right\rceil - 1 \).

**Proof.** For Statement 1, we generalize the class of binary partial words \( \{u_0, u_1, u_2, u_3\} \) used in the proof of Proposition 2. Given some \( n = 4q + r \) where \( 0 \leq r < 4 \), we can construct binary partial words \( u_{r,h} \) and \( u_{r,h+1} \) both of length \( n \) with \( h \) and \( h+1 \) holes, respectively, such that \( \text{sd}(u_{r,h+1}) - \text{sd}(u_{r,h}) \geq 2q - 1 \). Define

\[
\begin{align*}
  u_{0,h} &= \diamond \left\lceil \frac{h}{2} \right\rceil 0^q \left\lfloor \frac{h}{2} \right\rfloor^{-1} 10^q 2 110^q 10^q \left\lfloor \frac{h}{2} \right\rfloor^{1} \diamond \left\lceil \frac{h}{2} \right\rceil, \\
  u_{0,h+1} &= \diamond \left\lceil \frac{h}{2} \right\rceil 0^q \left\lfloor \frac{h}{2} \right\rfloor^{-1} 10^q 2 10^q \left\lfloor \frac{h}{2} \right\rfloor^{1} \diamond \left\lceil \frac{h}{2} \right\rceil.
\end{align*}
\]

The following are additional squares present in \( u_{0,h+1} \) but not in \( u_{0,h} \):

- \( q - 1 \) squares of the form \((0^s10^t)^2\) for \( s + t = q \) with \( s = 1, 2, \ldots, q - 1 \) requiring the new hole to be 0;

- \( q \) squares of the form \((0^s10^t)^2\) for \( s + t = q - 1 \) with \( s = 0, 1, \ldots, q - 1 \) requiring the new hole to be 0.

Hence \( \text{sd}(u_{0,h+1}) - \text{sd}(u_{0,h}) = 2q - 1 \).
Define
\[ u_{1,h} = \bigcirc \left\lfloor \frac{h}{2} \right\rfloor 0^q - \left\lceil \frac{h}{2} \right\rceil^{-1} 10^q - 100^q \bigcirc \left\lceil \frac{h}{2} \right\rceil , \]
\[ u_{1,h+1} = \bigcirc \left\lfloor \frac{h}{2} \right\rfloor 0^q - \left\lceil \frac{h}{2} \right\rceil^{-1} 10^q - 100^q \bigcirc \left\lceil \frac{h}{2} \right\rceil . \]

The following are additional squares present in \( u_{1,h+1} \) but not in \( u_{1,h} \):
- \( q \) squares of the form \( (0^s 10^t)^2 \) for \( s + t = q \) with \( s = 0, 1, \ldots, q - 1 \) requiring the new hole to be 1;
- \( q-1 \) squares of the form \( (0^s 10^t)^2 \) for \( s + t = q - 1 \) with \( s = 0, 1, \ldots, q - 2 \) requiring the new hole to be 1.

Note that the square \( (0^{q-1}1)^2 \) already exists as a suffix of \( u_{1,h} \), and hence \( sd(u_{1,h+1}) - sd(u_{1,h}) = 2q - 1 \).

Define
\[ u_{2,h} = \bigcirc \left\lfloor \frac{h}{2} \right\rfloor 0^q - \left\lceil \frac{h}{2} \right\rceil^{-1} 10^q - 100^q \bigcirc \left\lceil \frac{h}{2} \right\rceil + 1 , \]
\[ u_{2,h+1} = \bigcirc \left\lfloor \frac{h}{2} \right\rfloor 0^q - \left\lceil \frac{h}{2} \right\rceil^{-1} 10^q - 100^q \bigcirc \left\lceil \frac{h}{2} \right\rceil + 1 . \]

The following are additional squares present in \( u_{2,h+1} \) but not in \( u_{2,h} \):
- \( q \) squares of the form \( (0^s 10^t)^2 \) for \( s + t = q \) with \( s = 0, 1, \ldots, q - 1 \) requiring the new hole to be 1;
- \( q-1 \) squares of the form \( (0^s 10^t)^2 \) for \( s + t = q - 1 \) with \( s = 0, 1, \ldots, q - 3 \) and \( s = q - 1 \) requiring the new hole to be 1 if \( h > 2 \); \( q \) squares of the form \( (0^s 10^t)^2 \) for \( s + t = q - 1 \) with \( s = 0, 1, \ldots, q - 1 \) requiring the new hole to be 1 if \( h \leq 2 \).

Hence \( sd(u_{2,h+1}) - sd(u_{2,h}) \geq 2q - 1 \).

Define
\[ u_{3,h} = 01 \bigcirc \left\lfloor \frac{h}{2} \right\rfloor 0^q - \left\lceil \frac{h}{2} \right\rceil^{-1} 10^q - 100^q \bigcirc \left\lceil \frac{h}{2} \right\rceil , \]
\[ u_{3,h+1} = 01 \bigcirc \left\lfloor \frac{h}{2} \right\rfloor 0^q - \left\lceil \frac{h}{2} \right\rceil^{-1} 10^q - 100^q \bigcirc \left\lceil \frac{h}{2} \right\rceil . \]

The following are additional squares of \( u_{3,h+1} \) that are not in \( u_{3,h} \):
- \( q - 1 \) squares of the form \( (0^s 10^t)^2 \) for \( s + t = q \) with \( s = 1, 2, \ldots, q - 1 \) requiring the new hole to be 0;
• $q$ squares of the form $(0^{s}10^{t})^2$ for $s + t = q - 1$ with $s = 0, 1, \ldots, q - 1$ requiring the new hole to be 0;
• 1 square $(10^{q}10^{q-1})^2$ requiring the new hole to be 0;
• 1 square $(010^{q}10^{q-2})^2$ requiring the new hole to be 0.

Hence $sd(u_{3,h+1}) - sd(u_{3,h}) = 2q + 1$.

For Statement 2, consider the words

$u = 1^l\lceil\frac{n}{2}\rceil\lceil\frac{n}{2}\rceil - 1\lceil\frac{n}{2}\rceil\lceil\frac{n}{2}\rceil$,

$v = 1^l\lceil\frac{n}{2}\rceil\lceil\frac{n}{2}\rceil - 1\lceil\frac{n}{2}\rceil\lceil\frac{n}{2}\rceil$.

If $n$ is even, we observe that any completion of the square partial word

$pq = 1^l\lceil\frac{n}{2}\rceil + 1\lceil\frac{n}{2}\rceil - 1\lceil\frac{n}{2}\rceil\lceil\frac{n}{2}\rceil$,

where $p = 1^l\lceil\frac{n}{2}\rceil + 1\lceil\frac{n}{2}\rceil - 1$ and $q = 1^l\lceil\frac{n}{2}\rceil\lceil\frac{n}{2}\rceil$, is compatible with $u$, whereas $pq \not\uparrow v$, so no completion of $pq$ is compatible with $v$. Therefore, the number of squares present in $u$ but not in $v$ is bounded from below by $2^l - 1 = \|H(p \lor q)\|$. If $n$ is odd, consider the square partial word

$pq = 1^l\lceil\frac{n}{2}\rceil + 1\lceil\frac{n}{2}\rceil - 1\lceil\frac{n}{2}\rceil\lceil\frac{n}{2}\rceil - 1$.

If $h$ is even, let $p = 1^l\lceil\frac{n}{2}\rceil + 1\lceil\frac{n}{2}\rceil - 1$ and $q = 1^l\lceil\frac{n}{2}\rceil\lceil\frac{n}{2}\rceil - 2$, while if $h$ is odd, let $p = 1^l\lceil\frac{n}{2}\rceil + 1\lceil\frac{n}{2}\rceil - 2$ and $q = 1^l\lceil\frac{n}{2}\rceil\lceil\frac{n}{2}\rceil - 2$. In either case, we see that all of the completions of $pq$ are compatible with the prefix of $u$ of length $n - 1$, whereas none of its completions are compatible with any factor of $v$. Therefore, the number of squares present in $u$ but not in $v$ is bounded from below by $2^l - 1$, where $\lceil\frac{n}{2}\rceil - 1 = \|H(p \lor q)\|$.

From direct calculation, we see that the first lower bound in Proposition 4 is greater only when $h \leq 2 \log_2(4|\frac{n}{4}| - 1)$.

**Proposition 5.** The inequality $\Gamma_2(n) \geq n + \lceil\frac{n-6}{8}\rceil - 7$ holds.

**Proof.** For $n = 1, \ldots, 9$, a computer check confirms the inequality. Let $n \geq 10$ and let $w_i$ be the partial word of length $n$ defined as below, where $l = \frac{n-4}{4}$ for each $i \in \{0, 1, 2, 3\}$. For $w_0 = 0^{l-2}10^{l-2}0^l10^l$, we have the following distinct squares:

\begin{align*}
\end{align*}
• $\lfloor \frac{l}{2} \rfloor$ squares of the form $(0^p)^2$ with $p = 1, \ldots, \lfloor \frac{l}{2} \rfloor$ independent of the holes;

• $l - 1$ squares of the form $(0^p 10^q)^2$ for $p + q = l$ and $p = 1, \ldots, l - 1$ each requiring the holes to be 0 and 1 in order;

• 1 square $(10^l)^2$ requiring only the second hole to be 1;

• $l - \lfloor \frac{l}{2} \rfloor$ squares of the form $(0^p)^2$ with $p = \lfloor \frac{l}{2} \rfloor + 1, \ldots, l$ each requiring both holes to be 0, except that $(0^{\lfloor \frac{l}{2} \rfloor + 1})^2$ requires only the second hole to be 0 if $l$ is odd;

• $l - 2$ squares of the form $(0^p 10^q)^2$ for $p + q = l + 1$ and $p = 1, \ldots, l - 2$ each requiring the holes to be 1 and 0 in order;

• $l - 1$ squares of the form $(0^p 10^q)^2$ for $p + q = l - 1$ and $p = 0, \ldots, l - 2$ each requiring the holes to be 0 and 1 in order;

• $l - 2$ squares of the form $(0^p 10^q)^2$ for $p + q = l - 2$ and $p = 0, \ldots, l - 3$ each requiring the holes to be 1 and 0 in order;

• 1 square $(0^{l-1})^2$ requiring only the first hole to be 1;

• 1 square $1^2$ requiring both holes to be 1.

So we have a difference of $(5l - 3) - (\lfloor \frac{l}{2} \rfloor + 3) = n + \lceil \frac{n}{8} \rceil - 6$.

We similarly argue for $w_1 = 0^{l-2}10^{l-2}0\circ\circ0^l10^{l+1}$, $w_2 = 0^{l-1}10^{l-2}0\circ\circ0^l10^{l+1}$, and $w_3 = 0^{l-1}10^{l-1}0\circ\circ0^l10^l$.

Next, we prove the following facts about the behavior of $\Gamma_h(n)$ and $\Delta_h(n)$.

**Proposition 6.** The sequence $\{\Gamma_h(n)\}_{1 \leq h \leq n}$ for fixed $n$ is monotone increasing.

**Proof.** For any $1 \leq h < n$, find a partial word $w$ of length $n$ with $h$ holes and a completion $\hat{w}$ such that $sd(w) - sd(\hat{w}) = \Gamma_h(n)$. For some position $i \in D(w)$, obtain $w'$ from $w$ by replacing the letter in position $i$ with $\Diamond$, $D(w)$ is nonempty since $h < n$. Clearly $sd(w') \geq sd(w)$ and $w' \subset \hat{w}$. Thus we have $sd(w') - sd(\hat{w}) \geq sd(w) - sd(\hat{w}) = \Gamma_h(n)$. Therefore $\Gamma_{h+1}(n) \geq \Gamma_h(n)$, whenever $h < n$. \(\square\)
The sequence \( \{\Gamma_h(n)\}_{n \geq 1} \) for fixed \( h \) is not monotone increasing in general. For example, \( \Gamma_2(8) = 7 \) and \( \Gamma_2(9) = 6 \), which can be verified through direct computation.

A computer check shows that \( \Delta_5(5) = 1 \), while \( \odot\odot\odot\odot\odot \) and its strengthening \( \odot\odot\odot\odot \) give witness to \( \Delta_4(5) = 3 \). Therefore the sequence \( \{\Delta_h(n)\}_{1 \leq h \leq n} \)

is not monotone increasing for fixed \( n \).

### 4.2 Arbitrary alphabet

Let us now give a tight upper bound on \( sd_k(w, i) \), where \( w \) is a \( k \)-ary partial word with \( h \) holes. We start with a lemma whose proof is similar to that of [8, Theorem 2.1], which only deals with the case \( h = 1 \).

**Lemma 2.** If \( w \) is a partial word with \( h \) holes over a \( k \)-letter alphabet, then \( sd_k(w, i) \leq 2k^h \) for all \( 0 \leq i < |w| \).

**Proof.** From [7], for any full word \( w \), \( sd_k(w, i) \leq 2 \) for all \( 0 \leq i < |w| \). Use this as a base case and induct on \( h \). Assume our result holds for words with less than \( h \) holes. Now assume for the sake of contradiction that there exists a partial word \( w \) with \( h \) holes such that \( sd_k(w, i) > 2k^h \) for some \( 0 \leq i < |w| \).

By the pigeonhole principle, there exists some letter \( a \) in our alphabet so that, when replacing one of the holes by \( a \) to obtain the strengthening \( \tilde{w} \), we have \( sd_k(\tilde{w}, i) > 2k^{h-1} \). Since \( \|H(\tilde{w})]\| = h - 1 \), this contradicts the inductive hypothesis. \( \square \)

We now modify the construction from [8, Section 2] to construct a partial word \( w \) with \( h \) holes over \( k \) letters which witnesses the upper bound \( sd_k(w, i) = 2k^h \). To begin, recall Halava et al.’s construction: let \( w_0 = \odot a_{k-1}a_{k-2} \ldots a_0 \), and recursively define \( w_{2j-1} = w_{2j-2}w_{2j-2}(a_j-1) \) and \( w_{2j} = w_{2j-1}[\odot^{-1}w_{2j-1}a_j^{-1}] \) for \( j = 1, 2, \ldots, k^h \) where \( w_i(a_j) \) denotes the completion of \( w_i \) with all \( \odot \)'s set to \( a_j \). Here \( a_j^{-1} \) means we remove the suffix \( a_j^{-1} \) from \( w_{2j-1} \) and \( \odot^{-1} \) means we remove the prefix \( \odot \) from \( w_{2j-1} \). Using this construction, they showed that \( sd_k(w_{2k}, 0) = 2k \) [8, Theorem 2.5], and consequently, \( sd_{k^h}(w_{2k^h}, 0) = 2k^h \). Note that \( w_{2k^h} \) has one hole.

**Theorem 6.** There exists a word \( w \) over a \( k \)-letter alphabet with \( h \) holes such that \( sd_k(w, i) = 2k^h \).
Proof. We begin by constructing a word $w$ such that $\text{sd}(w, 0) = 2k^h$. Let $w_i$ as above. Define $\mathcal{A} = \{a_0, a_1, \ldots, a_{k^h-1}\}$, $\mathcal{B} = \{b_0, b_1, \ldots, b_{k-1}\}$, and $\mathcal{C} = \{u_0, u_1, \ldots, u_{k^h-1} : u_i$ is the $(i + 1)$-th word of length $h$ over $\mathcal{B}$ in lexicographical order}. We now define a morphism $\phi$ from $\mathcal{A}$ by $\phi(a_i) = b_0^h b_1 u_i$ and $\phi(\phi) = b_0^h b_1^h$. For example, if $h = 2$ and $k = 2$, $\phi(a_1) = b_0^2 b_1 b_0 b_1$.

Applying this morphism to $w_{2k^h}$ produces a word $w$ with $h$ holes over $k$ letters. This morphism induces an injective map $f$ from the square occurrences of $w_{2k^h}$ to the square occurrences of $w$. Let $S$ be the set of squares in $w_{2k^h}$ which have their rightmost occurrence beginning at position 0, and let $T = f(S)$ be the set of images of all the squares in $S$ under $f$. Every element of $T$ is a square occurrence which begins at position 0 in $w$. Also notice that $f : S \to T$ is a bijection. We claim that no square in $T$ can occur later in $w$. Suppose towards a contradiction that some square $u$ in $T$ has another occurrence $v$ later in $w$. There are two cases to consider.

Suppose $v$ is the image of a square occurrence in $w_{2k^h}$ which starts after position 0. Then $f^{-1}(u)$ occurs later in $w_{2k^h}$, which violates the definition of $S$.

Now, suppose $v$ is not the image of a square occurrence in $w_{2k^h}$. We know that $v$ begins with $b_0^h b_1$. Note that $b_0^h b_1$ is unbordered, i.e. no proper prefix is also a suffix, and also note that its length is $h + 1$. Therefore, the only occurrences of $b_0^h b_1$ in $w$ are as prefixes of $f(x)$ for some completion $x$ of a factor of $w_{2k^h}$. But then, $v$ must be the image of a square occurrence in $w_{2k^h}$, which is a contradiction.

Since every square in $T$ has its last occurrence at position 0 and $|T| = |S| = 2k^h$, $\text{sd}_k(w, 0) = 2k^h$.

Finally let $w' = b_0^h w$. Since $\text{sd}_k(w', i)$ only counts the last appearance of each distinct square, $\text{sd}_k(w', i) = 2k^h$. $\square$

The following two propositions give some necessary conditions for a $k$-ary partial word $w$ with $h$ holes to satisfy $\text{sd}_k(w, i) = 2k^h$ for some positions $i$.

**Proposition 7.** If a partial word $w$ with $h$ holes is such that $\text{sd}_k(w, i) = 2k^h$, then $j \in D(w)$ for all $0 \leq j < i$.

*Proof.* Assume there is a word $w$ with $h$ holes such that $\text{sd}_k(w, i) = 2k^h$, and $j \in H(w)$ yet $j < i$. If we strengthen $w$ at $j$, we obtain $\bar{w}$ with $h - 1$ holes, but since $j < i$, we have $\text{sd}_k(\bar{w}, i) = \text{sd}_k(w, i) = 2k^h$, a contradiction with Lemma 2. $\square$
**Proposition 8.** If a partial word $w$ with $h$ holes is such that $\text{sd}_k(w, i) = 2^k h$ for all $0 \leq i < m$, then $|w| > 2m$.

**Proof.** By [11], if a full word $w$ is such that $\text{sd}_k(w, i) = 2^k h$ for all $0 \leq i < m$, then $|w| > 2m$. Use this as a base case and induct on $h$. If $\text{sd}_k(w, i) = 2^k h$, then $\text{sd}_k(\bar{w}, i) = 2^{k-1} h$ for any strengthening $\bar{w}$ of $w$, and so $|w| = |\bar{w}| > 2m$.

The following proposition shows that for a word $w$ with $h$ holes and a position $i$ such that $\text{sd}_k(w, i) > 2^h$, the holes are restricted to certain squares.

**Proposition 9.** Let $w$ be a binary partial word with $h \geq 2$ holes, and suppose $\text{sd}_k(w, i) > 2^h$ for some $0 \leq i < |w|$, and denote these squares $u_1, u_2, \ldots, u_r$ with $|u_1| \leq |u_2| \leq \cdots \leq |u_r|$. Then there must be at least $j$ holes in $u_{2j}$ for $j = 1, \ldots, h$.

**Proof.** We know from Lemma 2 that the maximum number of distinct squares having their last occurrence at the same position in a binary word with $h$ holes is $2^{h+1}$.

For $j \geq 1$, suppose the $j$-th hole is not in $u_{2j}$. Since $\text{sd}_k(w, i) > 2^h \geq 2^j$, there is a way to fill the $j$-th, $(j+1)$-th, $\ldots$, $h$-th holes such that some square other than $u_1, \ldots, u_{2j}$ is at position $i$, and hence at least $2^j + 1$ distinct squares are at position $i$ with $j$ holes left. This contradicts Lemma 2.

To illustrate Proposition 9, consider the partial word


delimiter

The following squares occur uniquely at position 0 of $w$:

- $u_1 = \circ10010$,
- $u_2 = \circ10010\circ001$,
- $u_3 = \circ10010\circ0011\circ001$,
- $u_4 = \circ10010\circ0011\circ0011\circ0011\circ001$,
- $u_5 = w$.

There is at least one $\circ$ in $u_2$ and there are at least two $\circ$'s in $u_4$.

The next theorem gives an upper bound for the number of distinct squares in a partial word of a given length with one hole over an arbitrary
alphabet. Upper bounds are also given in [4, 8], but we give a much more detailed analysis, which yields a bound different from both of them. We will explain where the differences come from after the proof.

We require two lemmas.

**Lemma 3.** [5] If a partial word with one hole has at least three distinct squares at the same position, then the hole is in the shortest square.

**Lemma 4.** [4] If \(w^2, v^2\) and \(u^2\) are three square full words at the same position in a partial word with one hole, with \(|w| < |v| < |u|\), then \(2|w| \leq |u|\).

**Theorem 7.** The inequality \(sd_{1,k}(n) \leq (3 + \gamma)n + O(\log n)\), where \(\gamma = \sum_{i=2}^{\infty} \frac{1}{2^{i-1}} = 0.60669515 \cdots\) holds.

**Proof.** In the proof, we might omit the \(O(\log n)\) term for convenience when it does not affect the analysis. Denote the position of the hole as \(j\). We want to find an upper bound on the number of distinct squares using Lemma 3 and Lemma 4.

We first claim that \(\frac{j}{n} \leq \frac{1}{2}\) witnesses the upper bound found by this approach. Otherwise, suppose \(\frac{j}{n} > \frac{1}{2}\). Then if three distinct squares start at a position \(i \leq 2j - n + 1\), the shortest square must contain the hole by Lemma 3, and the third shortest square has length at least \(2(j - i + 1)\) by Lemma 4, but this is impossible since it would end no earlier than \(i + 2(j - i + 1) - 1 = 2j - i + 1 \geq n\). Therefore all positions no later than \(2j - n + 1\) start at most 2 squares by [7]. Ignoring this beginning part, the rest of the word can be estimated as the same problem with a smaller length \(n'\) and \(\frac{j'}{n'} = \frac{1}{2}\) where \(j'\) is the new hole position. As shown below, an upper bound on the number of distinct squares in this case is found to be \(cn'\) for some constant \(c > 2\) using Lemma 3 and Lemma 4. Then taking into account the ignored beginning part, the number of distinct squares will be less than \(cn\). But when \(\frac{j}{n} = \frac{1}{2}\), we get an upper bound \(cn'\) of distinct squares, and hence \(\frac{j}{n} \leq \frac{1}{2}\) witnesses the upper bound if we use this approach.

Now we compute an upper bound on the number of distinct squares starting at each position. Each position after \(j\) starts at most 2 distinct squares, and each position \(i \leq j\) could start either an odd or even number of distinct squares. The higher one should be taken as an upper bound.

Suppose \(2s + 1\) distinct squares start at \(i\) where \(s \geq 1\). We sort them by length and denote them by \(u_1, \ldots, u_{2s+1}\). By Lemma 3, \(u_1\) ends no earlier than \(j\), so \(|u_1| \geq j - i + 1\). By Lemma 4, \(|u_3| \geq 2|u_1|, |u_5| \geq 2|u_3|, \ldots, |u_{2s+1}| \geq 2|u_{2s-1}|\). Since \(|u_{2s+1}| \leq n - i, (j - i + 1)2^s \leq n - i, \) or \(2s + 1 \leq 2|\log_2(\frac{n-i}{j-i+1})| + 1\).
Now suppose $2s + 2$ distinct squares start at $i$ where $s \geq 1$. Again we sort them by length and denote them by $u_1, \ldots, u_{2s+2}$. Again $|u_1| \geq j - i + 1$. Since there is only one hole, $|u_2| \geq |u_1| + 2$. Also $|u_{2s+2}| \leq n - i$, so $(j - i + 3)2^s \leq n - i$, or $2s + 2 \leq 2[\log_2(\frac{n-i}{j-i+3})] + 2$.

To compare these two numbers, we first assume that an even number of distinct squares gives a higher estimate:

$$[\log_2 \left( \frac{n-i}{j-i+1} \right)] \leq [\log_2 \left( \frac{n-i}{j-i+3} \right)].$$  \hspace{1cm} (1)$$

Since $\frac{n-i}{j-i+1} > \frac{n-i}{j-i+3}$, (1) is violated if and only if

$$\log_2 \left( \frac{n-i}{j-i+1} \right) \geq m > \log_2 \left( \frac{n-i}{j-i+3} \right)$$

for some integer $m$. There are at most $O(\log n)$ different values that $m$ can take. To find at most how many different position $i$'s correspond to a certain $m$, denoted as $\Delta i$, we let

$$\frac{n-i}{j-i+1} = \frac{n-(i+\Delta i - 1)}{j-(i+\Delta i - 1) + 3}$$

where $i$ refers to the minimum value satisfying (2) for given $n, j, m$. Equation (3) can be solved for $\Delta i$ to give

$$\Delta i = 2 \frac{j-i+1}{n-j+1} + 3 = O(1)$$

since $\frac{j}{n} \leq \frac{1}{2}$. Therefore there are at most $O(\log n)$ position $i$'s where (1) is violated. The estimate using an odd number of distinct squares can exceed the estimate using an even number of distinct squares by at most $O(1)$, so by assuming that an even number of distinct squares start at each $i$, $i \leq j$, we get an error of $O(\log n)$ on the upper bound. Hence for each position $i$, $i \leq j$, we overestimate the number of distinct squares starting there, $2s + 2$, by letting $s = [\log_2(\frac{n-i}{j-i+3})]$ at position $i$. Then $s$ can take values from the lowest $[\log_2(\frac{n-j}{j+3})]$ to the highest $[\log_2(\frac{n-j}{j-1})]$. We denote the number of positions $i$, $i \leq j$ such that $s = p$ as $\beta(p)$. By definition, the sum of $\beta(p)$ over all possible $p$'s is $j + 1$. Then our upper bound estimate is

$$2(n-j-1) + \sum_{p=\lfloor \log_2(\frac{n-j}{j+3}) \rfloor}^{\lfloor \log_2(\frac{n-j}{j-1}) \rfloor} \beta(p)(2p + 2) = 2n + 2 \sum_{p=\lfloor \log_2(\frac{n-j}{j+3}) \rfloor}^{\lfloor \log_2(\frac{n-j}{j+3}) \rfloor} p\beta(p).$$  \hspace{1cm} (4)$$
Note that we will be omitting some floors in deriving formulas involving \( \beta \), but it will become clear that this changes the sum (7), that appears below, by \( O(\log n) \), which is negligible in our analysis. To obtain \( \beta(p) \), notice that \( s \leq p \) at position \( i \) if and only if
\[
\log_2 \frac{n}{j+3} - i \leq p < \frac{n-j-3}{2^{p+1} - 1}.
\] (5)

In particular, when \( p \) takes on the minimum index \( \lfloor \log_2 \frac{n}{j+3} \rfloor \),
\[
\beta(p) = \beta \left( \left\lfloor \log_2 \left( \frac{n}{j+3} \right) \right\rfloor \right) = j + 3 - \frac{n-j-3}{2^{\lfloor \log_2 \frac{n}{j+3} \rfloor + 1} - 1}.
\] (6)

Otherwise, for \( \lfloor \log_2 \frac{n}{j+3} \rfloor < p \leq \lfloor \log_2 \frac{n-j}{3} \rfloor \), we take the difference of (5) between two consecutive \( p \)'s,
\[
\beta(p) = \left( j + 3 - \frac{n-j-3}{2^{p+1} - 1} \right) - \left( j + 3 - \frac{n-j-3}{2^p - 1} \right) = (n-j-3) \left( \frac{1}{2^p - 1} - \frac{1}{2^{p+1} - 1} \right).
\]

Hence our upper bound estimate (4) becomes
\[
2n + 2(n-j-3) \sum_{p=\lfloor \log_2 \frac{n}{j+3} \rfloor}^{\lfloor \log_2 \frac{n-j}{3} \rfloor} \frac{1}{2^{p+1} - 1} + 2 \left\lfloor \log_2 \left( \frac{n}{j+3} \right) \right\rfloor \beta \left( \left\lfloor \log_2 \left( \frac{n}{j+3} \right) \right\rfloor \right) \beta \left( \left\lfloor \log_2 \left( \frac{n}{j+3} \right) \right\rfloor \right) \beta \left( \left\lfloor \log_2 \left( \frac{n}{j+3} \right) \right\rfloor \right).
\] (7)

We argued at the beginning that \( \frac{j}{n} \leq \frac{1}{2} \) witnesses the upper bound. Now we want to maximize (7). The sum in the second term of (7) equals
\[
\sum_{p=\lfloor \log_2 \frac{n}{j+3} \rfloor + 2}^{\lfloor \log_2 \frac{n-j}{3} \rfloor} \frac{1}{2^{p+1} - 1} + \frac{\lfloor \log_2 \frac{n-j}{3} \rfloor + 1}{2^{\lfloor \log_2 \frac{n-j}{3} \rfloor + 1} - 1} - \frac{\lfloor \log_2 \frac{n-j}{3} \rfloor}{2^{\lfloor \log_2 \frac{n-j}{3} \rfloor + 1} - 1} \quad \text{by telescoping.}
\] (8)

We plug (8) back into (7), and see that the contribution of the third term of (8) in (7) is
\[
2(n-j-3) \left( \frac{\lfloor \log_2 \frac{n-j}{3} \rfloor}{2^{\lfloor \log_2 \frac{n-j}{3} \rfloor + 1} - 1} \right).
\]
whose absolute value is bounded from above by

\[
2(n - j - 3) \frac{\lfloor \log_2 \left( \frac{n}{3} \right) \rfloor}{2^{\log_2 \left( \frac{n}{3} \right)} - 1} = 2(n - j - 3) \frac{\lfloor \log_2 \left( \frac{n}{3} \right) \rfloor}{n/3 - 1} = 6 \left\lfloor \log_2 \left( \frac{n - j}{3} \right) \right\rfloor = O(\log n),
\]
so we ignore it. By writing \( \lfloor \log_2 \left( \frac{n}{j+3} \right) \rfloor \) as \( q \) and ignoring the above mentioned term, (7) becomes

\[
2n + 2(n - j - 3) \left( \sum_{p=q+2}^{\lfloor \log_2 \left( \frac{n}{j+3} \right) \rfloor} \frac{1}{2^p - 1} + \frac{q + 1}{2^{q+1} - 1} \right) + 2q \beta(q). \tag{9}
\]

Furthermore, by defining \( \beta(p) = \alpha(p) - \alpha(p + 1) \), where \( \alpha(p) = (n - j - 3)/(2^p - 1) \) for all \( p \geq q \) (thus \( \alpha(q) = j + 3 \)), and after rearranging, (9) becomes

\[
2n + 2 \left( \sum_{p=q+1}^{\lfloor \log_2 \left( \frac{n}{j+3} \right) \rfloor} \alpha(p) \right) + 2q \alpha(q). \tag{10}
\]

If \( q = 0 \), which requires \( \frac{1}{2} - \frac{3}{n} < \frac{j}{n} \leq \frac{1}{2} \), then (9) equals

\[
2n + 2(n - j - 3) \left( \sum_{p=2}^{\lfloor \log_2 \left( \frac{n}{j+3} \right) \rfloor} \frac{1}{2^p - 1} + 1 \right), \tag{11}
\]

which is maximized by taking \( \frac{j}{n} = \frac{1}{2} - \frac{3}{n} \). We get

\[
2n + n \left( \sum_{p=2}^{\lfloor \log_2 \left( \frac{n}{j+6} \right) \rfloor} \frac{1}{2^p - 1} + 1 \right). \tag{12}
\]

If instead \( q = 1 \), which requires \( \frac{1}{4} - \frac{3}{n} < \frac{j}{n} \leq \frac{1}{2} - \frac{3}{n} \), we plug (6) in, and (9) equals

\[
\begin{align*}
2n + 2(n - j - 3) & \left( \sum_{p=3}^{\lfloor \log_2 \left( \frac{n}{j+3} \right) \rfloor} \frac{1}{2^p - 1} + \frac{2}{3} \right) + 2 \left( j + 3 - \frac{n - j - 3}{3} \right) \\
= 2n + 2(n - j - 3) & \left( \sum_{p=3}^{\lfloor \log_2 \left( \frac{n}{j+3} \right) \rfloor} \frac{1}{2^p - 1} + \frac{2}{3} \right) + \frac{2}{3} (4j + 12 - n). \tag{13}
\end{align*}
\]
One can verify that by taking $\frac{j}{n} = \frac{1}{2} - \frac{3}{n}$ in (13), we get the same result as (12) in terms of the coefficient of $n$. We define the constant

$$\gamma = \sum_{i=2}^{\infty} \frac{1}{2^i - 1} = 0.60669515 \ldots.$$  

Then the finite sums in (12) and (13) approach $\gamma$ from below as $n \to \infty$.

To maximize (13) with respect to $j$, we take its derivative against $j$, which is at least

$$-2 \left( \gamma - \frac{1}{3} + \frac{2}{3} \right) + \frac{8}{3} > 0$$

where $\frac{1}{3}$ comes from the fact that the first sum over $p$ in (13) is missing the term $\frac{1}{p-1}$ for $p = 2$ compared to the definition of $\gamma$. So in fact (13) is increasing as $j$ increases. Hence (12) gives at least as high an estimate as (13).

In general, for an integer $q > 1$, the term $2q\alpha(q)$ of (10) can be bounded by $n$ and twice the finite sum by $\gamma n$, actually by $(\gamma - \frac{1}{3})n$, where the $\frac{1}{3}$ term arises for the same reason as above, with some error term of $O(\log n)$.

We conclude that an upper bound on the number of distinct squares is

$$(3 + \gamma)n + O(\log n) \approx 3.6067n + O(\log n).$$

We briefly make some comments on the upper bounds $3.5n$, obtained in [4], and $4n$, obtained in [8], for the number of distinct squares in a partial word of length $n$ with one hole over an arbitrary alphabet. Our bound, approximately $3.61n$, is weaker than $3.5n$. The $4n$ bound basically uses the same idea as the one presented here, but uses a weaker $2\log_2((n-i)/(j-i+1)) + 2$ bound than our bound

$$\max(2\log_2((n-i)/(j-i+3)) + 2, 2\log_2((n-i)/(j-i+1)) + 1)$$

for the number of distinct squares at position $i$. Neither of [4, 8] discusses the possibility of having an odd or even number of distinct squares starting at a position before or at the hole; [4] assumes an odd number throughout and [8] assumes an even number throughout. Both [4, 8] use integrals to estimate sums of discrete values. Since a significant portion of the summands take small integer values, errors are created in such estimates (resulting in either underestimates or overestimates). It is our much more detailed analysis, based on sums instead of integrals, which makes the differences between the earlier bounds and our bound.
5 Square Occurrences

A square occurrence at position $i$ in a full word is any occurrence of a factor $x^2$ starting at position $i$. In partial words, for each factor $xy$ with $x \uparrow y$, we count every full word $z^2$ such that $xy \subset z^2$ as a square occurrence. Given such factor $xy$, the number of square occurrences that $xy$ contributes to is $k\|H(xy)\|$, where $k$ is the alphabet size. Note that square occurrences can potentially overlap and they need not be distinct.

We use $so_k(w)$ to denote the number of square occurrences in a $k$-ary partial word $w$, and we let

$$so_{h,k}(n) = \min\{so_k(w) : w \in A^*_n, |w| = n, \|H(w)\| = h\},$$

where $A$ is any alphabet of size $k$. To simplify the notation, we will abbreviate $so_k(w)$ by $so(w)$ and $so_{h,2}(n)$ by $so_h(n)$.

5.1 Binary alphabet

We find the limit of the ratio $\frac{so_h(n)}{n}$.

Borrowing from [12], we construct a partial word that achieves the minimal number of square occurrences in the limit for a given number of holes. Let

$$w_{X,Y} = 01001101000110010110001101001100011001100101100011001010011010001100101001110010110011101001101011001011100110100.$$

The word $w_{X,Y}$ has length 187. Let $w_a = w_{0,0\overline{0},0}, w_b = w_{0,0\overline{1},0},$ and $w_c = w_{1,0\overline{1},1}$. Let $w_\circ$ be the partial word with one hole in position 0 of $w_a$ and the rest of the letters the same as $w_a$. Define the morphism $g : \{a,b,c\}^* \rightarrow \{0,1\}^*$ by $g(a) = w_a, g(b) = w_b,$ and $g(c) = w_c$. Let $t$ be a word over the alphabet $\{a,b,c\}$ and denote $g(t)$ by $t'$. Let $t'_\circ$ be $g(t)$ with the first $h$ occurrences of $w_a$ replaced by $w_\circ$. If $x$ is a factor of $t$, we call $x'$ the corresponding factor of $t'$ and $x'_\circ$ the corresponding factor of $t'_\circ$.

Lemma 5. For a ternary square-free word $t$, each square occurrence of $t'_\circ$ belongs to exactly one of the following three categories:

1. It is completely inside $w_a, w_\circ, w_b,$ or $w_c$;
2. It is one of the squares crossing the boundary of adjacent blocks of the form \( w_\alpha w_\diamond \) where \( \alpha \in \{a, \diamond, b, c\} \): 11, 1010, 0101, 1001110011, and 1100111001, compatible with

\[
1 \diamond, 1010, 0101, 100110011, \text{ and } 110011001,
\]

respectively;

3. It is either 0101 or 1010 crossing the boundary of adjacent blocks of the form \( w_\alpha w_\beta \) where \( \alpha \in \{a, \diamond, b, c\} \), \( \beta \in \{a, b, c\} \).

Thus \[\text{so}(t'_c) = \begin{cases} 
206|t| - 2, & h = 0; \\
206|t| + 4h - 5, & h > 0 \text{ and } a \text{ is the first letter of } t; \\
206|t| + 4h - 2, & \text{otherwise.}
\end{cases}\]

Proof. We modify the proof from [12] for partial words. Consider a ternary square-free word \( t \). Suppose there exists a factor of \( t'_c \) compatible with a full square \( q^2 \), i.e., not one of the above. Assume \( |q^2| < 4 \times 374 \). Then there exists a subword \( z \) of \( t \) with \( |z| \leq 5 \) such that \( q^2 \) is compatible with a factor of \( z'_c \). By a computer check, for each ternary square-free word \( z \) of length at most 5, \( z'_c \) has only the aforementioned square occurrences.

Now assume \( |q^2| \geq 4 \times 374 \). Say \( q^2 \) is compatible with the factor \( uv \) of \( t'_c \) where \( u \uparrow v \). Since \( |u| \geq 2 \times 374 \), \( u \) has at least one of \( w_a, w_\diamond, w_b, \) or \( w_c \) as a factor. Say that \( u \) has the factor \( w_\alpha \) where \( \alpha \in \{a, \diamond, b, c\} \). Then there exists \( w_\beta \), where \( \beta \in \{a, \diamond, b, c\} \), such that \( w_\beta \) is compatible with \( w_\alpha \) and \( w_\beta \) is \( |u| \) positions after \( w_\alpha \). A computer check shows that for \( x, y \in \{a, \diamond, b, c\} \), \( w_xw_y \) does not have any of \( w_a, w_\diamond, w_b, \) or \( w_c \) as a proper factor. If \( |u| \) is not a multiple of 374, then \( w_\beta \) would be a proper factor of \( w_xw_y \) for some \( x, y \in \{a, \diamond, b, c\} \). So \( |u| = 374l \) for some integer \( l \). If \( uv \) is centered at the boundary of two adjacent letters of \( t \), then \( t \) would have a square. So \( uv \) cannot be centered between adjacent letters.

Let \( x \) be the largest subword of \( t \) such that \( x'_c \) is entirely inside \( u \) and \( y \) be the subword of \( t \) with \( y'_c \) entirely inside \( v \) and \( y'_c \) is compatible with \( x'_c \). Observe that \( w_a, w_\diamond, w_b, \) and \( w_c \) are incompatible in exactly positions 108, 295, and 306. In those positions, \( w_a \) and \( w_\diamond \) has 0, 1, and 1; \( w_b \) has 0, 0, and 1; and \( w_c \) has 1, 0, and 0. Since the hole can only occur in the first position of \( w_\diamond \), \( x = y \). Since \( t \) is square-free, the smallest subword of \( t \) that allows \( t'_c \) to have a square must have at least a letter of \( t \) between the occurrences
of $x$. So the minimal subword $r$ of $t$ such that $r'$ has the factor $q^2$ is of the form $\alpha x\beta x\gamma$ with $|x| = l - 1$ and $\alpha, \beta, \gamma \in \{a, b, c\}$.

If the center of $q^2$ occurs before position 295 of $\beta'_o$, then the letters in positions 295 and 306 of $\alpha'_o$ are the same as those in $\beta'_o$. But $\alpha'_o$ cannot match $\beta'_o$ in those positions unless $\alpha = \beta$. This produces the square $(\alpha x)^2$ in $t$, which contradicts the assumption that $t$ is square-free. If the center of $q^2$ is after position 295, then positions 108 and 295 of $\beta'_o$ should match those in $\gamma'_o$, which would produce the square $(x\beta)^2$ in $t$, contradicting that $t$ is square-free. Hence $t'_o$ has only the squares mentioned in the lemma.

By a computer check, there are 204 square occurrences in $w_a$, $w_b$, and $w_c$ while $w_\diamond$ has 205 square occurrences. Across the boundary of $w_\alpha w_\beta$ for $\alpha \in \{a, \diamond, b, c\}$ and $\beta \in \{a, b, c\}$, there are 2 square occurrences, namely 0101 and 1010. For $\alpha \in \{a, \diamond, b, c\}$, the boundary of $w_\alpha w_\diamond$ has 5 square occurrences, namely 11, 1010, 0101, 1001110011, and 1100111001 compatible with 1\diamond, 1\diamond10, 0\diamond1, 1001\diamond1001, and 11001\diamond1001 respectively. If $h = 0$, then there are 204 square occurrences inside each of the $|t|$ images and 2 square occurrences crossing the $|t| - 1$ boundaries of adjacent letters. If $h > 0$ and the first letter of $t$ is $a$, then there are 204 square occurrences inside the $|t| - h$ images that do not have holes, 205 square occurrences in the $h$ appearances of $w_\diamond$, 5 square occurrences across the $h - 1$ boundaries of pairs of images that end with $w_\diamond$, and 2 square occurrences in the remaining $|t| - h$ boundaries. If $h > 0$ and the first letter of $t$ is not $a$, then there is one more boundary with 5 squares crossing it and one less boundary with 2 squares crossing it.

**Theorem 8.** Let $\{h_n\}$ be an integer sequence. If $\lim_{n \to \infty} \frac{h_n}{n} = 0$, then

$$\lim_{n \to \infty} \frac{\text{so}_{h_n}(n)}{n} = \frac{103}{187} \approx 0.55080.$$  

Moreover, if $\lim_{n \to \infty} \frac{h_n}{n} \leq \frac{15}{1343221}$, then

$$0.55080 \approx \frac{103}{187} \leq \lim_{n \to \infty} \frac{\text{so}_{h_n}(n)}{n} \leq \frac{739864}{1343221} \approx 0.55081.$$  

**Proof.** Note that $\text{so}_{h_n}(n) \geq \text{so}_0(n)$, so the lower bound for $\text{so}_{h_n}(n)/n$ follows from [15].

First, assume $\lim_{n \to \infty} \frac{h_n}{n} = 0$. From [13], there exists a ternary square-free word, which has infinitely many occurrences of $a$. Using a ternary square-free word $t$ of length $n$, let $t'_o$ be as above with the first $h_n$ occurrences of $w_a$ replaced by $w_\diamond$, so that the length of $t'_o$ is $374|t| = 374n$ and $t'_o$ has
$h_n$ holes. By Lemma 5, $206n - 2 \leq \mathsf{so}(t'_0) \leq 206n + 4h_n - 2$. Dividing each term by $374n$ and taking the limit $n \to \infty$, we have

$$
\frac{103}{187} \leq \lim_{n \to \infty} \frac{\mathsf{so}_n(t'_0)}{374n} \leq \frac{103}{187} + \frac{4}{374} \lim_{n \to \infty} \frac{h_n}{n} = \frac{103}{187},
$$

so the limit holds for the ratio $\frac{\mathsf{so}_n(374n)}{374n}$.

Now, assume $\lim_{n \to \infty} \frac{h_n}{n} \leq \frac{15}{13366}$. A result from [14] states that the maximal frequency of a letter in a square-free ternary word is $\frac{255}{653}$. Without loss of generality, suppose $t$ is a square-free ternary word such that $a$ occurs with this frequency. Let $t'_0$ as above, so that the first $h_n$ occurrences of $w_a$ are replaced by $w_0$. Then, the ratio of holes to length is $(\frac{255}{653})(\frac{1}{374}) = \frac{15}{13366}$. Using the same inequality as above, we obtain

$$
\frac{103}{187} \leq \lim_{n \to \infty} \frac{\mathsf{so}_n(t'_0)}{374n} \leq \frac{103}{187} + \frac{4}{374} \lim_{n \to \infty} \frac{h_n}{n} = \frac{739864}{1343221}.
$$

Above, $t'_0$ can only have length a multiple of 374. We slightly modify $t'_0$ to construct a binary partial word of any length. For $\alpha \in \{a, b, c\}$, consider $t'_0 w_\alpha[0..l-1]$, a partial word of length $374|t| + l = 374n + l$ with $h_n$ holes. From Lemma 5, $206|t| - 2 \leq \mathsf{so}(t'_0 w_\alpha[0..l-1]) \leq 206|t| + 4h_n + 204$, so in particular, appending $w_\alpha$ only contributes a constant number of square occurrences. Therefore, for $0 \leq l < 374$ the limit still holds when looking at the ratio $\frac{\mathsf{so}_n(374n+l)}{374n+l}$.

\section{5.2 Arbitrary alphabet}

We prove a result about the asymptotic behavior of the minimum ratio of square occurrences in $k$-ary partial words to length.

\begin{proposition}
Let $k \geq 3$. Given $\varepsilon > 0$, let $\{h_n\}$ be an integer sequence such that $\lim_{n \to \infty} \frac{h_n}{\log_k n} < 1 - \varepsilon$. Then

$$
\lim_{n \to \infty} \frac{\mathsf{so}_{h_n,k}(n)}{n} = 0.
$$

\end{proposition}

\begin{proof}
From [13], there exists a $k$-ary square-free word for all $k \geq 3$, which we denote by $w'$. By Lemma 2, the maximum number of distinct squares having their rightmost occurrence at a position in a partial word with $h_n$ holes over a $k$-letter alphabet is $2k^{h_n}$.

Consider $w = \diamond w'$. Since no position of $w'$ starts a square and prepending a $\diamond$ does not affect the square-freeness of $w'$, any square occurrence at
position 0 in $w$, i.e. any full square compatible with a factor of $w$ beginning at position 0, is a distinct square occurrence. Therefore, $s_{o_k}(w) = sd_k(w,0) \le 2k$.

Now, consider $w = \omega^{h_n}w'$, with $h_n$ as in the statement of the proposition, and consider the number of square occurrences at $i$ for each $0 \le i < h_n$. At each $i$, $sd_k(w,i) \le 2k^{h_n-i}$ since $w[i..|w|]$ contains $h_n - i$ $\omega$’s. Additionally, there can be at most the number of distinct squares with their rightmost occurrences at positions $i+1, i+2, \ldots, h_n - 1$, or at most $\sum_{j=i+1}^{h_n-1} 2k^{h_n-j}$ square occurrences at position $i$. Therefore,

$$s_{o_k}(\omega^{h_n}w') \le \sum_{i=0}^{h_n-1} \sum_{j=i+1}^{h_n-1} 2k^{h_n-j} + \sum_{i=0}^{h_n-1} 2k^{h_n-i} \le 2\frac{k(k^{h_n+1} - 1)}{(k-1)^2} - 2(h_n + 1)\frac{k}{k-1}.$$ 

Thus, if $\lim_{n \to \infty} \frac{h_n}{\log_k n} < 1 - \varepsilon$,

$$\lim_{n \to \infty} \frac{s_{o_k}(\omega^{h_n}w')}{n} \le \lim_{n \to \infty} \frac{C_n^{1-\varepsilon}}{n} = 0$$

for some constant $C$. Since $s_{o_{h_n,k}}(n)$ is the minimum number of square occurrences, $\lim_{n \to \infty} \frac{s_{o_{h_n,k}}(n)}{n} = 0$. \hfill \boxrule{0.1cm}

### 6 Conclusion

For future work, we list some conjectures for the binary alphabet.

**Conjecture 1.** For all $n \ge 0$, $\Delta_1(n) \approx \frac{n}{2}$.

For all $8 \le n \le 25$, $\Delta_1(n) = \left\lfloor \frac{n-1}{2} \right\rfloor$, but $\Delta_1(26) = 13$. The only witness (up to reversal and complement) of length 26 which exhibits a difference of 13 is 001010100101 $\omega^{1010010101001}$.

**Conjecture 2.** For all $n \ge 8$, $\Delta_2(n) = \left\lfloor \frac{7(n-1)}{10} \right\rfloor$.

Conjecture 2 has been verified up to $n = 21$.

**Conjecture 3.** The equality $s_{o_1}(n) = s_{o_0}(n) + 1$ holds for all $n \ge 2$.

We now discuss Conjecture 3 in some details. We first prove the following proposition.
Proposition 11. The inequality $s_0(n) \geq s_0(n) + 1$ holds for all $n \geq 2$ and $0 \leq h < n$.

Proof. Notice that all the square occurrences in a partial word remain after we replace a letter with a hole. In particular, $s_0(n) \geq s_0(n)$.

We prove the result by contradiction. Suppose $s_0(n) = s_0(n)$ for some $n \geq 2$, and $w$ is a witness to $s_0(n)$. We choose a hole in $w$, and define $w_0$ and $w_1$ as $w$ with this hole filled by 0, 1, respectively. Then by the observation above, for $i = 0, 1$, $s_0(w_i) \leq s_0(w)$.

On the other hand, since $s_0(w) = s_0(n)$, by definition, $s_0(w) \geq s_0(w)$. Hence $s_0(w_0) = s_0(w_1) = s_0(n) = s_0(n) = s_0(w)$. As $n \geq 2$, the hole in $w$ has at least one neighbor, which is compatible with some letter $a \in B$. Then the square occurrence $aa$ comprising this neighbor and the hole appears in $w$ but not in $w_0$. This means $s_0(w) > s_0(w_0)$, a contradiction. \[\square\]

Hence Conjecture 3 is reduced to “$s_0(n) \leq s_0(n) + 1$ for all $n \geq 2”$.

A searching algorithm for determining $s_0(n)$ is described in [12], which finds, for each $n$ and each prefix of length 6, the full word that minimizes square occurrences. The computed data then can be used to prune branches in later searches. Kucherov et al. computed $s_0(n)$ up to $n = 3,300$. However, even with the pruning, since there are $2^n$ words of length $n$, the algorithm takes a very long time when $n$ is large.

We make an improvement on the algorithm to speed up the search, and also modify it to find $s_0(n)$. Consequently, we are able to verify Conjecture 3 up to $n = 4,000$ by computer. Note that it is not always true that placing a hole in position 0 of a witness for $s_0(n)$ gives a witness for $s_0(n)$. For $n = 16$, $s_0(101100101110010) = s_0(16)$ but $s_0(101100101110010) > s_0(16)$.

Let $T$ be a set of positive integers. Define $s_0^T(w)$ to be the number of square occurrences in $w$ that have root-lengths in $T$, and define

$$s_0^T(w) = \min\{s_0^T(w) : w \in B^*_\xi, |w| = n, ||H(w)|| = h\}.$$

In particular, $s_0^T(n) = s_0(n)$ for any $n, h$. We have the following proposition.

Proposition 12. Let $w$ be a witness to $s_0^T(n)$ for some $n, h, T$. If all the square occurrences in $w$ have root-lengths in $T$, then $w$ is also a witness to $s_0^T(n)$, and $s_0^T(n) = s_0^T(n)$.

Proof. Because $s_0(w) = s_0^T(w)$ for any word $u$, we have $s_0^T(n) = s_0^T(n)$. Also $s_0^T(n) \leq s_0(w)$ by definition. But since $s_0(w) = s_0^T(w) = s_0^T(n)$, $s_0^T(n) = s_0^T(n) = s_0(w)$. \[\square\]
The algorithm in [12] can be easily modified to compute $\mathcal{s}_h^T(n)$. The only change is that instead of checking all root-lengths, we only check the root-lengths in $T$. If the size of $T$ is small compared to $n$, the search will be much faster. Also if it happens that a witness to $\mathcal{s}_h^T(n)$ has all square occurrences with root-lengths in $T$, then by Proposition 12, we obtain $\mathcal{s}_h(n)$. So the problem is to choose a suitable $T$. We do this iteratively.

Given $h, N$, we compute $\mathcal{s}_h(n)$ for all $n \leq N$. Initially, let $T$ be some small set, for example $\{1, 2, 3, 4, 5, 6, 7\}$. We run the algorithm above. If for some $n$, the witness to $\mathcal{s}_h^T(n)$ that we find has square occurrences with root-lengths not in $T$, we add those root-lengths to $T$ and restart the search from $n$. It turns out that the size of $T$ grows very slowly as $n$ increases. For example, we are able to compute $\mathcal{s}_0(n)$ and $\mathcal{s}_1(n)$ up to $n = 4,000$ with a set $T$ of size about 20 (instead of considering all 2,000 possible root-lengths).

This gives us some insight into the problem in general. It appears that to minimize the number of square occurrences in a word, it suffices to minimize the number of square occurrences with some specific root-lengths.

We also suggest a number of open problems for an arbitrary alphabet. Referring to Proposition 7, which positions and how many positions can achieve $2k^h$? We know $\mathcal{s}_d(w, i) < 2k^h$ if there exists $j \in H(w)$ for some $0 \leq j < i$. How does this affect the maximum number of distinct squares in partial words? Ilie [11] limited the number of positions $i$ for which $\mathcal{s}_d(w, i) = 2$ in full words by looking at maximum runs of consecutive 2’s. This allowed for the $\Theta(\log n)$ improvement. Can similar approaches be used with partial words?

In addition, a World Wide Web server interface has been established at

www.uncg.edu/cmp/research/squares3

for automated use of a program that given as input a binary partial word over the alphabet $\{0, 1\}$, outputs the number of positions that start squares and the number of square occurrences. It also outputs the positions that start each square occurrence and the corresponding squares for each position.

Acknowledgements

We thank the referees of preliminary versions of this paper for their very valuable comments and suggestions. We thank Justin Lazarow for the proof of Lemma 1. We also thank Michelle Bodnar for her help in clarifying ideas in this paper.
References


