Acta Cybernetica **19** (2009) 465–477.

Counting Distinct Squares in Partial Words*

F. Blanchet-Sadri†, Robert Mercaş‡, and Geoffrey Scott§

Abstract

A well known result of Fraenkel and Simpson states that the number of distinct squares in a word of length n is bounded by $2n$ since at each position there are at most two distinct squares whose last occurrence start. In this paper, we investigate the problem of counting distinct squares in partial words, or sequences over a finite alphabet that may have some "do not know" symbols or "holes" (a (full) word is just a partial word without holes). A square in a partial word over a given alphabet has the form uu' where u' is *compatible* with u, and consequently, such square is compatible with a number of full words over the alphabet that are squares. We consider the number of distinct full squares compatible with factors in a partial word with h holes of length n over a k-letter alphabet, and show that this number increases polynomially with respect to k in contrast with full words, and give bounds in a number of cases. For partial words with one hole, it turns out that there may be more than two squares that have their last occurrence starting at the same position. We prove that if such is the case, then the hole is in the shortest square. We also construct a partial word with one hole over a k-letter alphabet that has more than k squares whose last occurrence start at position zero.

Keywords: combinatorics on words, partial words, squares

1 Introduction

Computing repetitions such as squares in sequences or strings of symbols from a finite alphabet is profoundly connected to numerous fields such as biology, computer

*This material is based upon work supported by the National Science Foundation under Grant No. DMS–0452020. This work was done during the second author's stay at the University of North Carolina at Greensboro. A World Wide Web site has been created at `www.uncg.edu/cmp/research/freeness` for this research.

†Department of Computer Science, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402–6170, USA, E-mail: `blanchet@uncg.edu`

‡GRLMC, Universitat Rovira i Virgili, Plaça Imperial Tárraco, 1, Tarragona, 43005, Spain and MOCALC Research Group, Faculty of Mathematics and Computer Science, University of Bucharest, Academiei, 14, 010014, Bucharest, Romania

§Department of Mathematics, Dartmouth College, 6188 Kemeny Hall, Hanover, NH 03755–3551, USA

science, and mathematics [8]. The stimulus for recent works on repetitions in strings is the study of biological sequences such as DNA that play a central role in molecular biology. In addition to its sheer quantity, repetitive DNA is striking for the variety of repetitions it contains, for the various proposed mechanisms explaining the origin and maintenance of repetitions, and for the biological functions that some of the repetitions may play. The literature has generally considered problems in which a period u of a repetition is invariant. It has been required that occurrences of u match each other exactly. In some applications however, such as DNA sequence analysis, it becomes interesting to relax this condition and to recognize u' as an occurrence of u if u' is *compatible* with u.

A well known result of Fraenkel and Simpson [3] states that the number of distinct squares in a word of length n is bounded by $2n$ since at each position there are at most two distinct squares whose last occurrence start. In [6], Ilie improves this bound to $2n - \Theta(\log n)$. Based on numerical evidence, it has been conjectured that this number is actually less than n. In this paper, we investigate the problem of counting distinct squares in partial words, or sequences over a finite alphabet that may contain some "do not know" symbols or "holes." In Section 2, after making some remarks about the maximum number of distinct full squares compatible with factors of a partial word, we give some lower bounds for that number. These bounds are related to the length of the word, the alphabet size this word is defined on, and the number of holes it contains. In Section 3, we show that for partial words with one hole, there may be more than two squares that have their last occurrence starting at the same position. We prove that if such is the case, then the hole is in the shortest square. There, we also construct for $k \geq 2$, a partial word with one hole over a k-letter alphabet that has more than k squares whose last occurrence start at position 0. Finally in Section 4, we provide some conclusions and suggestions for future work.

We end this section by reviewing basic concepts on partial words. Fixing a nonempty finite set of letters or an *alphabet* A, a *partial word* u of length $|u| = n$ over A is a partial function $u : \{0, \ldots, n-1\} \to A$. For $0 \leq i < n$, if $u(i)$ is defined, then i belongs to the *domain* of u, denoted by $i \in D(u)$, otherwise i belongs to the *set of holes* of u, denoted by $i \in H(u)$. The unique word of length 0, denoted by ε, is called the *empty* word. For convenience, we will refer to a partial word over A as a word over the enlarged alphabet $A_\diamond = A \cup \{\diamond\}$, where $\diamond \notin A$ represents a hole. The set of all words (respectively, partial words) over A of finite length is denoted by A^* (respectively, $A_\diamond^*$).

The partial word u is *contained in* the partial word v, denoted by $u \subset v$, provided that $|u| = |v|$, all elements in $D(u)$ are in $D(v)$, and for all $i \in D(u)$ we have that $u(i) = v(i)$. As a weaker notion, u and v are *compatible*, denoted by $u \uparrow v$, provided that there exists a partial word w such that $u \subset w$ and $v \subset w$. An equivalent formulation of compatibility is that $|u| = |v|$ and for all $i \in D(u) \cap D(v)$ we have that $u(i) = v(i)$. We denote by $u \vee v$ the least upper bound of u and v, that is, for every partial word w such that $u \subset w$ and $v \subset w$, we have $(u \vee v) \subset w$. If $u \not\uparrow v$, then we adopt the convention that $u \vee v = \varepsilon$. The following rules are useful for computing with partial words: (1) *Multiplication:* If $u \uparrow v$ and $x \uparrow y$,

then $ux \uparrow vy$; (2) *Simplification:* If $ux \uparrow vy$ and $|u| = |v|$, then $u \uparrow v$ and $x \uparrow y$; and (3) *Weakening:* If $u \uparrow v$ and $w \subset u$, then $w \uparrow v$.

A partial word u is *primitive* if there exists no word v such that $u \subset v^n$ with $n \geq 2$. If u is a nonempty partial word, then there exist a primitive word v and a positive integer n such that $u \subset v^n$. Uniqueness holds for full words but not for partial words as seen with $u = \diamond a$ where $u \subset a^2$ and $u \subset ba$ for distinct letters a, b. For partial words u, v, w, if $w = uv$, then u is a *prefix* of w, denoted by $u \leq w$, and if $v \neq \varepsilon$, then u is a *proper prefix* of w, denoted by $u < w$. If $w = xuy$, then u is a *factor* of w. If $u = u_1u_2$ for some nonempty compatible partial words u_1 and u_2, then u is called a *square*. Whenever we refer to a square u_1u_2 it will imply that $u_1 \uparrow u_2$.

2 Counting distinct squares: A first approach

In a full word, every factor of length $2n$ contains at most one square factor ww with $|w| = n$. In a square partial word w_0w_1 where $w_0 \uparrow w_1$, we call the word $v = w_0 \vee w_1$ the *general form* of the square. For example, the general form of the square $ab\diamond\diamond c\diamond a\diamond d\diamond\diamond\diamond$ is $abd\diamond c\diamond$. We observe that in partial words, a square w_0w_1 may be compatible with more than one distinct full square of length $2|w_0|$. For example, the word $aa\diamond aa\diamond$ over the alphabet $\{a, b, c\}$ is compatible with three distinct full squares of length 6: $(aaa)^2$, $(aab)^2$ and $(aac)^2$. It is easy to see that if $aa\diamond aa\diamond$ is a word over an alphabet of size k, then it is compatible with exactly k squares of length 6. Whenever we talk about a full square compatible with a general form, we refer to a square that has the first half compatible with the general form. In general, if $w = a_0a_1 \dots a_{2m-1}$ is a partial word over a k-letter alphabet A, and w is a square, then w is compatible with exactly $k^{\|H(v)\|}$ squared full words of length m, where $v = a_0a_1 \dots a_{m-1} \vee a_ma_{m+1} \dots a_{2m-1}$.

At this point, we see that the study of distinct squares in partial words is quite different from the study of distinct squares in full words. In the case of full words, there exists an upper bound for the number of distinct squares in a word of length n, no matter what the alphabet size is. The same statement is certainly untrue for partial words. For example, the number of distinct nonempty full squares compatible with $\diamond\diamond$ is equal to k, where k is the alphabet size.

Let w be a partial word over a k-letter alphabet A. We will denote by $f_k(w)$ the number of distinct nonempty full squares over A compatible with factors of w, and by $g_{h,k}(n)$ the maximum of the $f_k(w)$'s where w ranges over all partial words of length n with h holes, over alphabet A. Note that the number of all distinct full square nonempty words compatible with factors of $\diamond^n$, where n is a positive integer, over A, is equal to the number of all distinct full nonempty words of length $i \leq \lfloor \frac{n}{2} \rfloor$ over A. Using this remark,

$$g_{n,k}(n) = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} k^i = \frac{k\ (k^{\lfloor \frac{n}{2} \rfloor} - 1)}{k-1} \tag{1}$$

Note that if n is odd, then $g_{n-1,k}(n-1) = g_{n,k}(n)$ and $g_{n-1,k}(n) = g_{n,k}(n)$. The first equality follows directly from (1). For the second equality, note that the number of distinct nonempty full squares compatible with factors of $\diamond^{n-1}a$ over the k-letter alphabet A where $a \in A$ is at least $g_{n-1,k}(n-1) = g_{n,k}(n)$ (those compatible with factors of $\diamond^{n-1}$). Thus, $g_{n-1,k}(n) \geq g_{n,k}(n)$. Since the function $g_{h,k}(n)$ is clearly monotonically increasing with respect to h, k, and n, it follows that $g_{n-1,k}(n) \leq g_{n,k}(n)$. Thus, $g_{n-1,k}(n) = g_{n,k}(n)$.

As we have seen earlier with the word $\diamond\diamond$, the number of distinct nonempty full squares compatible with factors of a partial word may be unbounded if we allow the alphabet size to grow arbitrarily large. However, we can often write this number as a function of the alphabet size. The following proposition shows that this number is indeed a polynomial in the alphabet size.

Proposition 1. *Let w be a partial word of length n over a k-letter alphabet, and let S_1 be the set of general forms of all factors of w that are squares. Let S_m be the set of all partial words v that can be written as $v = u_0 \vee u_1 \vee \cdots \vee u_{m-1}$, where $u_i \in S_1$ for all $0 \leq i < m$ and $u_i \neq u_j$ for all $i < j < m$. Then the number of full distinct squares compatible with factors of w is given by*

$$\sum_{m=1}^{\lfloor \frac{n}{2} \rfloor} ((-1)^{m-1} \sum_{s \in S_m} k^{\|H(s)\|}) \tag{2}$$

Proof. For a set X of partial words, denote by $\hat{X}$ the set of all full words compatible with elements of X. The number of full distinct square words compatible with factors of w is given by $\|\hat{S_1}\|$. By the principle of inclusion-exclusion,

$$\hat{S_1} = \sum_{m=1}^{\lfloor \frac{n}{2} \rfloor} ((-1)^{m-1} \sum_{s \in S_m} \|\{\hat{s}\}\|)$$

Since $\|\{\hat{s}\}\| = k^{\|H(s)\|}$, the proof is complete. □

To generalize the study of counting distinct squares in words to partial words, we are interested in the limit behaviour of $g_{h,k}(n)$ as k increases. However, as we have seen with the word $w = \diamond\diamond$, the value $\lim_{k\to\infty} f_k(w)$ may be infinity. Following Proposition 1, if we treat k as an unknown variable, the number of distinct nonempty full squares compatible with factors in any partial word is a polynomial with respect to k. If we consider all such polynomials corresponding to words of length n containing h holes, the maximal such polynomial would describe this limiting behavior. Given a finite length n, there exist only finitely many partial words of length n up to an isomorphism between letters. Therefore, a lower bound for $g_{h,k}(n)$ can be given using the leading term of this well defined maximal polynomial, $m_{h,k}(n)$.

The next results give bounds on the leading term in $m_{h,k}(n)$. We begin by defining a *free hole* of a square. Let w be a partial word over an alphabet A that

contains a factor v that is a square. A hole in v is called a *free hole* of v if the square v is preserved even after we replace the hole with any letter of A. For example, consider the partial word $w = ab\diamond a\diamond\underline{\diamond}$ over the alphabet $\{a, b, c\}$. The underlined hole is a free hole of the squares $ab\diamond a\diamond\diamond$ and $\diamond\diamond$, but not of $\diamond a\diamond\diamond$. It is easy to see that the number of free holes of a square factor is exactly twice the number of holes in the general form of that square. Two free holes in positions i and j in a square v are aligned if $i = j + \frac{|v|}{2}$ or $j = i + \frac{|v|}{2}$ and $v(i) = v(j) = \diamond$.

Note that the degree of $m_{h,k}(n)$ is $\lfloor \frac{h}{2} \rfloor$. To see this, let w be a word of length n with h holes over a k-letter alphabet. Clearly, any factor of w that is a square has at most $\lfloor \frac{h}{2} \rfloor$ holes in its general form. Thus, by (2) there can be no term of $m_{h,k}(n)$ with k raised to a power higher than $\lfloor \frac{h}{2} \rfloor$. Also note that the word $w = \diamond^h a^{n-h}$ achieves this bound. The following technical lemma will assist us in proving results about the coefficients of $m_{h,k}(n)$.

Lemma 1. *Let l be a positive integer, let w be a partial word of length n, and let $0 \leq p_1 \leq p_2 < n$. Then there are at most $\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1$ factors $v = w(i)w(i+1)\dots w(i+2l-1)$ of length $2l$ in w such that $i \leq p_1$ and $i + l > p_2$.*

Proof. Assume that there exist $\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 2$ such factors of length $2l$ in w. Since all of these factors have the same length, no two of them may start at the same position. Therefore, $p_1 \geq \lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1$. In particular, one of these factors must start at a position no later than $p_1 - (\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1)$. This gives us that $l > ((p_2 - p_1) + \lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1)$ from the condition that $i + l > p_2$. For any factor $v = w(i)w(i+1)\dots w(i+2l-1)$ of length $2l$ in w, we know that the length of w must exceed $2l + i$. Since there exist $\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 2$ such factors, at least one must start at a position i satisfying $i \geq \lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1$. Therefore, we obtain the contradiction

$$n \geq 2(p_2 - p_1 + \lfloor \frac{n - 2(p_2 - p_1 + 1)}{3} \rfloor + 2) + \lfloor \frac{n - 2(p_2 - p_1 + 1)}{3} \rfloor + 1$$

$$n \geq 3\lfloor \frac{n - 2(p_2 - p_1 + 1)}{3} \rfloor + 2(p_2 - p_1 + 1) + 3$$

$$n \geq n - 2(p_2 - p_1 + 1) - 2 + 2(p_2 - p_1 + 1) + 3$$

□

Intuitively, the above lemma states that for any $l > 0$, there can be at most $\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1$ factors of length $2l$ that use the letters $w(p_1)w(p_1+1)\dots w(p_2)$ in their first half. We will use this lemma to find upper bounds for the leading term of $m_{h,k}(n)$.

Theorem 1. *The leading term in $m_{2h,k}(n)$ is $(\lfloor \frac{n-2h}{3} \rfloor + 1)k^h$.*

Proof. The degree of $m_{2h,k}(n)$ being h, it only remains to show that the coefficient of k^h in $m_{2h,k}(n)$ is equal to $\lfloor \frac{n-2h}{3} \rfloor + 1$. We will give a lower bound of this

coefficient by constructing a word with the given leading term. Consider any word w of length n containing $2h$ holes and the factor

$$a^{\lfloor \frac{n-2h}{3} \rfloor} \diamond^h a^{\lfloor \frac{n-2h}{3} \rfloor} \diamond^h a^{\lfloor \frac{n-2h}{3} \rfloor}$$

The following is an exhaustive list of general forms of factors of w that are squares containing $2h$ free holes:

$$\begin{array}{ccccc} aaa & \ldots & aa\diamond\diamond & \ldots & \diamond\diamond \\ aaa & \ldots & a\diamond\diamond\diamond & \ldots & \diamond a \\ & & \vdots & & \\ a\diamond\diamond & \ldots & \diamond\diamond aa & \ldots & aa \\ \diamond\diamond\diamond & \ldots & \diamond aaa & \ldots & aa \end{array}$$

These $\lfloor \frac{n-2h}{3} \rfloor + 1$ partial words are pairwise compatible, but for any words v_1, v_2 in the above list, $\|H(v_1 \vee v_2)\| < h$. Therefore, by (2) we see that the coefficient of k^h in $m_{2h,k}(n)$ will be at least $\lfloor \frac{n-2h}{3} \rfloor + 1$.

Note that the coefficient of k^h corresponding to a word w is equal to the number of distinct factors in w, that are squares with $2h$ free holes. Let

$$w = w_0 \diamond_0 w_1 \diamond_1 w_2 \diamond_2 \ldots \diamond_{2h-1} w_{2h}$$

where $w_i \in A^*$ for all $0 \leq i \leq 2h$ and $\diamond_i = \diamond$ for all $0 \leq i < 2h$. Note that all factors of w with $2h$ free holes that are squares must have the same length (because in a square the free hole $\diamond_0$ is aligned with $\diamond_h$, the length of all such square factors will be twice the distance between $\diamond_0$ and $\diamond_h$). We observe that all factors of w that are squares containing $2h$ free holes must contain the first h holes of w in their first half. Therefore, every such factor contains $\diamond_0 w_1 \diamond_1 \ldots \diamond_{h-1}$ in its first half. The length of $\diamond_0 w_1 \diamond_1 \ldots \diamond_{h-1}$ is at least h, so by Lemma 1, there exist at most $\lfloor \frac{n-2h}{3} \rfloor + 1$ such factors. □

Proposition 2. *The leading term in $m_{2h+1,k}(n)$ is at least $(2\lfloor \frac{n-2h}{3} \rfloor + 1)k^h$.*

Proof. The degree of $m_{2h+1,k}(n)$ being h, it only remains to show that the coefficient of k^h in $m_{2h+1,k}(n)$ is at least $2\lfloor \frac{n-2h}{3} \rfloor + 1$. Consider any word w of length n containing $2h+1$ holes and the factor

$$a^{\lfloor \frac{n-2h}{3} \rfloor} \diamond^h a^{\lfloor \frac{n-2h}{3} \rfloor - 1} \diamond^{h+1} a^{\lfloor \frac{n-2h}{3} \rfloor}$$

The following is an exhaustive list of general forms of factors of w that are squares containing $2h$ free holes:

$$\begin{array}{ll} a^{\lfloor \frac{n-2h}{3} \rfloor - 1} a \diamond^{h-1} \diamond & a^{\lfloor \frac{n-2h}{3} \rfloor - 2} a \diamond^{h-1} \diamond \\ a^{\lfloor \frac{n-2h}{3} \rfloor - 1} \diamond\diamond^{h-1} a & a^{\lfloor \frac{n-2h}{3} \rfloor - 2} \diamond\diamond^{h-1} a \\ \vdots & \vdots \\ a \diamond^{h-1} \diamond a^{\lfloor \frac{n-2h}{3} \rfloor - 1} & \diamond^{h-1} \diamond a a^{\lfloor \frac{n-2h}{3} \rfloor - 2} \\ \diamond^{h-1} \diamond a^{\lfloor \frac{n-2h}{3} \rfloor - 1} a & \end{array}$$

There are $\lfloor\frac{n-2h}{3}\rfloor + 1$ words in the left column and $\lfloor\frac{n-2h}{3}\rfloor$ words in the right column. It is easy to check that if we select two compatible words v_1, v_2 from the above list of $(2\lfloor\frac{n-2h}{3}\rfloor + 1)$ partial words, $\|H(v_1 \vee v_2)\| < h$. Using (2) we get that the coefficient of k^h in $m_{2h+1,k}(n)$ will be at least $2\lfloor\frac{n-2h}{3}\rfloor + 1$. □

Proposition 3. *The leading term in $m_{2h+1,k}(n)$ is at most $(2\lfloor\frac{n-2h}{3}\rfloor + 3)k^h$ for $h > 1$.*

Proof. Let w be a word of length n containing $2h+1$ holes for some $h > 1$. Then w is of the form $w_0\diamond_0 w_1\diamond_1 w_2\diamond_2 \ldots \diamond_{2h} w_{2h+1}$ where $\diamond_i = \diamond$ for all i. We need to count the number of distinct factors of w that are squares containing $2h$ free holes. Let S denote the set of all such factors in w. Note that for every $s \in S$, there exists a hole in w that is not a free hole of s. Let S_j denote the set of all $s \in S$ having the property that $\diamond_j$ is not a free hole of s. Clearly, we have the partition $S = \cup_{0\le j\le 2h} S_j$.

First, assume that there exists $j \notin \{0, h, 2h\}$ such that $S_j \neq \emptyset$. Then $w_j\diamond_j w_{j+1} \uparrow w_k$ for some $j \neq k$. If there exists an i distinct from j such that $S_i \neq \emptyset$, then in one of the squares of S_i, the hole $\diamond_j$ is aligned with $\diamond_{k-1}$ or $\diamond_k$. In these cases, we get that $|w_{j+1}| \geq |w_k|$ or $|w_j| \geq |w_k|$ respectively. Both cases contradict with $w_j\diamond_j w_{j+1} \uparrow w_k$. Thus, $S_i = \emptyset$ for all $i \neq j$. Hence, we can replace $w_j\diamond w_{j+1}$ in w with w_k and preserve all squares. The resulting word has only $2h$ holes. From Theorem 1,

$$\|S\| \leq \lfloor\frac{n-2h}{3}\rfloor + 1$$

Next, let us consider the case where $S_j = \emptyset$ for every $j \notin \{0, h, 2h\}$. Note that all squares in S_0 have length equal to the distance between $\diamond_1$ and $\diamond_{h+1}$ in w, since these two holes are aligned in each square of S_0. Using the same argument, all squares in S_{2h} have length equal to the distance between $\diamond_1$ and $\diamond_{h+1}$ in w. Therefore, the length of squares in S_0 is equal to the length of the squares in S_{2h}. Note that all squares in S_0 and S_{2h} contain the factor $\diamond_1 w_2\diamond_2 \ldots \diamond_{h-1}$ in their first half. The length of this common factor is at least $h-1$. By Lemma 1, $\|S_0 \cup S_{2h}\| \leq \lfloor\frac{n-2(h-1)}{3}\rfloor + 1 = \lfloor\frac{n-2h+5}{3}\rfloor$. Since all squares in S_h have the same length and contain the factor $\diamond_0 w_1\diamond_1 \ldots \diamond_{h-1}$, it follows from Lemma 1 that $\|S_h\| \leq \lfloor\frac{n-2h}{3}\rfloor + 1$. Therefore,

$$\|S\| \leq \lfloor\frac{n-2h}{3}\rfloor + 1 + \lfloor\frac{n-2h+5}{3}\rfloor \leq 2\lfloor\frac{n-2h}{3}\rfloor + 3$$

The upper bound for $\|S\|$ reached in the second case is always greater than or equal to the upper bound reached in the first case. Therefore,

$$\|S\| \leq 2\lfloor\frac{n-2h}{3}\rfloor + 3$$

□

Proposition 4. *The leading term in $m_{3,k}(n)$ is at most $\frac{3n}{4}k$.*

Proof. Let $w = w_0 \diamond w_1 \diamond w_2 \diamond w_3$ be a partial word of length n with three holes. We wish to count the number of possible factors of w that are squares containing two free holes. Let S_1 be all such factors wherein the first hole of w is *not* free. Define S_2 and S_3 similarly. We wish to find the size of $S = \cup_{1 \leq i \leq 3} S_i$. The types of factors in S_1, S_2, and S_3 are illustrated below (the first half of each factor is written above the second half to show the alignment of the holes):

S_1	$(w_0 \diamond w_1)''$	$\diamond$	w_2'
	w_2''	$\diamond$	w_3'
S_2	w_0''	$\diamond$	$(w_1 \diamond w_2)'$
	$(w_1 \diamond w_2)''$	$\diamond$	w_3'
S_3	w_0''	$\diamond$	w_1'
	w_1''	$\diamond$	$(w_2 \diamond w_3)'$

where v' and v'' denote a prefix and suffix of a word v respectively. Because all factors in S_1 have the second and third holes of w aligned, all factors in S_1 have the same length. Therefore, each factor in S_1 ends at a different position of $\diamond w_3$. Also, the first element of the second half of each factor in S_1 occurs at a different position of $w_2 \diamond$. Therefore, $\|S_1\| \leq |w_3| + 1$ and $\|S_1\| \leq |w_2| + 1$. We can use similar reasoning to arrive at the following relations:

$$\|S_1\| \leq |w_2| + 1 \qquad \|S_2\| \leq |w_0| + 1 \qquad \|S_3\| \leq |w_0| + 1$$
$$\|S_1\| \leq |w_3| + 1 \qquad \|S_2\| \leq |w_3| + 1 \qquad \|S_3\| \leq |w_1| + 1$$

Because $\|S\| = \|S_1\| + \|S_2\| + \|S_3\|$ and $n = |w_0| + |w_1| + |w_2| + |w_3| + 3$, we determine that

$$\|S\| \leq |w_2| + 1 + |w_3| + 1 + |w_1| + 1 = n - |w_0|$$
$$\|S\| \leq |w_2| + 1 + |w_3| + 1 + |w_0| + 1 = n - |w_1|$$
$$\|S\| \leq |w_3| + 1 + |w_0| + 1 + |w_1| + 1 = n - |w_2|$$
$$\|S\| \leq |w_2| + 1 + |w_0| + 1 + |w_1| + 1 = n - |w_3|$$

Therefore,

$$\|S\| \leq n - \max\{|w_0|, |w_1|, |w_2|, |w_3|\} \leq n - \lceil \frac{n-3}{4} \rceil \leq \frac{3n}{4}$$

□

As we show next, we can improve the bound for the case when there are only two holes present in the word.

Proposition 5. *If* $n \equiv 2 \bmod 6$, *then*

$$m_{2,k}(n) - \frac{n+1}{3}k \geq \frac{n-2}{2}$$

Proof. Using Theorem 1 and the fact that $n \equiv 2 \bmod 6$, the leading term in $m_{2,k}(n)$ is $\frac{n+1}{3}k$. Therefore, $m_{2,k}(n) - \frac{n+1}{3}k$ is the constant term of the polynomial $m_{2,k}(n)$. It suffices to construct a partial word w with two holes over a k-letter alphabet A with $|w| = n \equiv 2 \bmod 6$ such that w contains $\frac{n+1}{3}k + \frac{n-2}{2}$ distinct squares. Consider the word

$$w = (ab)^l \diamond (ab)^l \diamond (ab)^l$$

of length n over A, such that a, b are distinct letters of A with $l = \frac{n-2}{6}$. The following is an exhaustive list of general forms of factors of w that are squares:

$$\begin{array}{llll}
(ab)^l\diamond, & b(ab)^{l-1}\diamond a, & \ldots, & \diamond(ab)^l \\
ab, & (ab)^2, & \ldots, & (ab)^{\lfloor \frac{l}{2} \rfloor} \\
ba, & (ba)^2, & \ldots, & (ba)^{\lceil \frac{l}{2} \rceil} \\
(ab)^0a, & (ab)^1a, & \ldots, & (ab)^{l-1}a \\
(ba)^0b, & (ba)^1b, & \ldots, & (ba)^{l-1}b
\end{array}$$

Figure 1 illustrates these squares for $n = 32$. These general forms are pairwise incompatible. Thus, there are a total of

$$(2l+1)k + \lfloor \frac{l}{2} \rfloor + \lceil \frac{l}{2} \rceil + l + l = (\frac{n-2}{3} + 1)k + 3l = \frac{n+1}{3}k + \frac{n-2}{2}$$

distinct full words that are squares compatible with factors of w. □

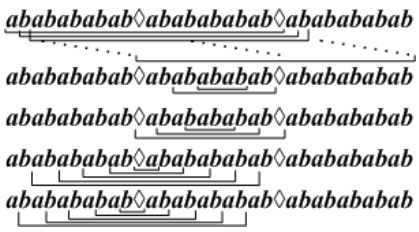

Figure 1: Squares in $(ab)^5\diamond(ab)^5\diamond(ab)^5$

3 Counting distinct squares: A second approach

At each position in a full word there are at most two distinct squares whose last occurrence starts, and thus the number of distinct squares in a word of length n is bounded by $2n$ as stated in the following theorem.

Theorem 2. [4] *Any full word of length n has at most $2n$ distinct squares.*

A short proof of Theorem 2 is given in [5]. It follows from the unique decomposition of words into primitive ones, and synchronization (a word w is primitive if and only if in ww there exist exactly two factors equal to w, namely the prefix and the suffix).

We now consider the one-hole case which behaves very differently from the zero-hole case. We will also count each square at the position where its last occurrence starts. If the last occurrence of a square in a partial word starts at position i, then it is a *square at position i*. In the case of partial words with one hole, there may be more than two squares that have their last occurrence starting at the same position. Such is the case with $a\diamond aababaab$ that has three squares at position 0: $a\diamond aa$, $a\diamond aaba$ and $a\diamond aababaab$. We will prove that if there are more than two squares at some position, then the hole is in the shortest square. We will also construct for $k \geq 2$, a partial word with one hole over a k-letter alphabet that has more than k squares at position 0. But first, we recall some results that will be useful for our purposes.

Lemma 2. [1] *Let $x, y \in A_\diamond^*$ be such that xy has at most one hole. If $xy \uparrow yx$, then there exist $z \in A^*$ and integers m, n such that $x \subset z^m$ and $y \subset z^n$.*

Lemma 3. [6] *Let $w \in A^*$. If $w = z_1z_2z_3 = z_2z_3z_4 = z_3z_4z_5$ for some $z_i \in A^* \setminus \{\varepsilon\}$, then there exist $x \in A^*$ primitive and integers p, q and r, $1 \leq p \leq r < q$, such that $x = x'x''$ for some $x' \in A^*$ and $x'' \in A^* \setminus \{\varepsilon\}$, and $z_1 = x^p$, $z_2 = x^{q-r}$, $z_3 = x^{r-p}x'$, $z_4 = x''x^{p-1}x'$, and $z_5 = x''x^{q-r-1}x'$.*

Theorem 3. *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.*

Proof. Let uu', vv' and ww' be the three shortest squares whose last occurrence start at the same position, and assume that $|w| < |v| < |u|$. It is impossible for these three squares to be all full (otherwise the subword u^2, a full word, would have three squares starting at its position 0).

For a contradiction, let us assume that ww' is full (here $w = w'$). If $w^2 \leq u$, then the prefix of length $|w^2|$ of u' is a later occurrence of a square compatible with w^2. And so we must have $v < u < w^2$. If the hole is in u' but not in v', then $v = v'$, and by replacing the hole with the corresponding letter in u, we obtain the full word u^2 that has three distinct squares at position 0, a contradiction. If the hole is in v', then set $w^2 = uz_3$, $u = vz_2$ and $v = wz_1$. We get $w = z_1z_2z_3$, $v = z_1z_2z_3z_1$ and $u = z_1z_2z_3z_1z_2$. Let w_2 and w_3 be the prefixes of length $|w|$ of v' and u' respectively. Since z_2z_3 is a prefix of both v and v', let z_4 be such that $w, w_2 \subset z_2z_3z_4$. Note that $|z_4| = |z_1|$. Two cases occur.

Case 1. The hole is in the suffix of length $|v| - |w|$ of v'.

In this case, let z_5 be such that $w = z_3z_4z_5$. Note that $|z_5| = |z_2|$. Here $w = z_1z_2z_3 = z_2z_3z_4 = z_3z_4z_5$ and by Lemma 3, there exist $x \in A^*$ primitive and integers p, q and r, $1 \leq p \leq r < q$, such that $x = x'x''$ for some $x' \in A^*$, $x'' \in A^* \setminus \{\varepsilon\}$, and $z_1 = x^p$, $z_2 = x^{q-r}$, $z_3 = x^{r-p}x'$, $z_4 = x''x^{p-1}x'$, and $z_5 = x''x^{q-r-1}x'$.

We have $w = z_1z_2z_3 = x^qx'$, $v = wz_1 = x^qx'x^p$ and $u = vz_2 = x^qx'x^px^{q-r}$. If $x' = \varepsilon$, then a later occurrence of a square compatible with w^2 exists, and so we assume that $x' \neq \varepsilon$. Since the hole is in the suffix of length $|v| - |w|$ of v', the hole is in the suffix of length $|x^p|$ of v'. We can write $v' = x^qx'x^sx_1x_2x^{p-s-1}$ where $0 \leq s < p$, $|x_1| = |x'|$ and $|x_2| = |x''|$, and where the hole is in x_1 or x_2. Since $u \uparrow u'$, we have $z_1z_2z_3z_1z_2 \uparrow z_3z_4x^sx_1x_2x^{p-s-1}\ldots$, or $x^qx'x^px^{q-r} \uparrow x^rx'x^sx_1x_2x^{p-s-1}\ldots$. The fact that $r < q$ implies that $x^{q-r}x'x^px^{q-r} \uparrow x'x^sx_1x_2x^{p-s-1}\ldots$. If $s > 0$, then $x'x''x' = x'x'x''$ and $x''x' = x'x''$, and the latter being an equation of commutativity implies that a word y exists such that $x' = y^m$ and $x'' = y^n$ for some integers m, n. In this case, there is obviously a later occurrence of a square compatible with w^2. If $s = 0$, then $x^{q-r}x'x^px^{q-r} \uparrow x'x_1x_2x^{p-1}\ldots$. Since $q > r$, by looking at the prefixes of length $|xx'|$ we get $x'x''x' \uparrow x'x_1x_2$ and deduce $x''x' \uparrow x_1x_2$.

If the hole is in x_1, then $x_2 = x''$ and $x''x' \uparrow x_1x''$. By weakening, we get $x''x_1 \uparrow x_1x''$, an equation of commutativity that satisfies the conditions of Lemma 2 since $x''x_1$ has only one hole. Similarly as above, a word y exists such that $x_1 \subset y^m$ and $x'' = y^n$ for some integers m, n. Set $x_1 = y^ty'y^{m-t-1}$ where $0 \leq t < m$ and y' is the factor that contains the hole. Since $x_1 \subset x'$, we deduce that $x' = y^ty''y^{m-t-1}$ for some y''. The compatibility $x''x' \uparrow x_1x''$ implies $y^ny^ty''y^{m-t-1} \uparrow y^ty'y^{m-t-1}y^n$ and by simplification $y^ny'' \uparrow y'y^n$. Since $x'' \neq \varepsilon$, we have $n > 0$ and obtain $y'' = y$. We get $x' = y^m$, and there is obviously a later occurrence of a square compatible with w^2. We argue similarly in the case where the hole is in x_2.

Case 2. The hole is not in the suffix of length $|v| - |w|$ of v'.

In this case, set $w = z_2z_3z_4$ and $w_2 = z_2z_3z_4'$ and the hole is in z_4'. Also, set $w = z_3z_4''z_5$ and $w_3 = z_3z_4'z_5$ where both $z_4' \subset z_4$ and $z_4' \subset z_4''$, and $|z_5| = |z_2|$. We treat the case where $z_4'' \neq z_4$ and leave the case where $z_4'' = z_4$ to the reader.

If $z_4'' \neq z_4$, then put $z_1 = x^p$ where x is primitive and p is a positive integer. Since $z_1z_2z_3 = z_2z_3z_4$ and the equation $z_1(z_1z_2z_3) = (z_1z_2z_3)z_4$ is one of conjugacy, we can write $z_4 = x''x^{p-1}x'$, where $x = x'x''$ with x'' nonempty, and $z_1z_2z_3 = x^qx'$ for some $q \geq p$. Since $z_1z_2z_3 = x^qx'$ and $z_1 = x^p$, we have $z_2z_3 = x^{q-p}x'$. Say $z_2 = x^ty'$ where $t \geq 0$, and y' is a prefix of x with $y' \neq x$. Set $x = y'y''$ with y'' nonempty. If $y' = \varepsilon$, we have $z_2 = x^t$ and $z_3 = x^{q-p-t}x'$ and in this case $z_4'' = z_4$, a contradiction. This can be seen by using the equality $z_2z_3z_4 = z_3z_4''z_5$. And so $y' \neq \varepsilon$. Since z_4' has the length of z_1, write $z_4' = (x''x')^sx_2x_1(x''x')^{p-s-1}$ where $0 \leq s < p$, $|x_1| = |x'|$, $|x_2| = |x''|$, and where the hole is in x_1 or x_2. There are three cases to consider: (2.1) $t < q-p-1$; (2.2) $t = q-p-1$; and (2.3) $t = q-p$. We prove the second one, and leave the other two to the reader.

For (2.2), $z_2 = x^ty'$ and $z_3 = y''x'$. Since $z_1z_2z_3 = z_3z_4''z_5$, we have $x^qx' = y''x'\ldots$. We consider the case where $|x'| \geq |y'|$ and then the case where $|x'| < |y'|$. If $|x'| \geq |y'|$ or y' is a prefix of x', then since $q = p + t + 1 > 0$, the prefixes of length $|x|$ are $y'y''$ and $y''y'$ respectively and again, the equality $y'y'' = y''y'$ holds, and as above leads to a contradiction. If $|x'| < |y'|$ or x' is a prefix of y', then since $z_1z_2z_3 \uparrow z_3z_4'z_5$, we have $x^qx' \uparrow y''x'(x''x')^sx_2x_1(x''x')^{p-s-1}\ldots$.

If $s > 0$, then the fact that the prefixes of length $|x|$ are compatible implies that $y'y'' = y''y'$. If $s = 0$ and the hole is in x_1, then $x_2 = x''$ and $y''x'x'' = y''x = y''y'y''$ is a prefix of $z_3z_4'z_5$ in which case $y'y'' = y''y'$ as above. If $s = 0$ and the hole is in x_2, then $x_1 = x'$ and set $y' = x'y$ for some $y \neq \varepsilon$. Here, $x'' = yy''$, and put $x_2 = y_1y_2$ where $y_1 \subset y$ and $y_2 \subset y''$. We get $x^qx' \uparrow y''x'x_2x_1(x''x')^{p-1} \ldots = y''x'y_1y_2x'(x''x')^{p-1} \ldots$.

If the hole is in y_2, then $y_1 = y$ and $y''x'y_1 = y''x'y = y''y'$ is a prefix of $z_3z_4'z_5$ and the result again follows since $y'y'' = y''y'$. If the hole is in y_1, then $y'y'' \uparrow y''x'y_1$ or $x'yy'' \uparrow y''x'y_1$, and by weakening $(x'y_1)y'' \uparrow y''(x'y_1)$. The latter being an equation of commutativity, by Lemma 2, we get that $x'y_1 \subset z^m$ and $y'' = z^n$ for some word z and positive integers m, n. Set $x'y_1 = z^kz'z^{m-k-1}$ where $0 \leq k < m$ and z' is the factor that contains the hole. Since $x'y_1 \subset x'y$, we deduce that $x'y = z^kz''z^{m-k-1}$ for some z''. The compatibility $x'yy'' \uparrow y''x'y_1$ implies $z^kz''z^{m-k-1}z^n \uparrow z^nz^kz'z^{m-k-1}$. By simplification we obtain $z''z^n \uparrow z^nz'$, and since $n > 0$ we get $z'' = z$, and thus $y' = x'y = z^m$. The result follows since $x = y'y'' = z^{m+n}$ with $m + n > 1$. □

Proposition 6. *For $k \geq 2$, there exists a partial word with one hole over a k-letter alphabet that has more than k squares at position 0.*

Proof. Let $\Sigma = \{a_1, a_2, \ldots\}$ be an infinite ordered set. We build a sequence of partial words with one hole, $(DS_i)_{i\geq 2}$, where DS_i contains $i + 1$ squares with their last occurrence starting at position 0. In order to do this, we build an intermediary sequence of partial words with one hole $(DS_i')_{i\geq 2}$ and denote by $DS_i'(a)$, the word DS_i' in which the hole has been replaced by the letter a. Let $DS_2 = a_1\diamond a_1a_1a_2a_1a_2a_1a_1a_2$, and for $i \geq 3$,

$$\begin{aligned} DS_{i-1}' &= DS_{i-1}a_{i-1} \\ DS_i &= DS_{i-1}'DS_{i-1}'(a_i) \end{aligned}$$

In other words, DS_i consists of the concatenation of DS_{i-1} with the last letter of the smallest alphabet used for creating DS_{i-1}, concatenated again with the same factor in which the hole has been replaced by a letter not present in the word so far. For example,

$$\begin{aligned} DS_2' &= a_1\diamond a_1a_1a_2a_1a_2a_1a_1a_2a_2 \\ DS_3 &= a_1\diamond a_1a_1a_2a_1a_2a_1a_1a_2a_2a_1a_3a_1a_1a_2a_1a_2a_1a_1a_2a_2 \end{aligned}$$

the latter having three squares other than itself at position 0: $a_1\diamond a_1a_1a_2a_1a_2a_1a_1a_2$, $a_1\diamond a_1a_1$ and $a_1\diamond a_1a_1a_2a_1$. For $k \geq 2$, DS_k, a partial word with one hole over a k-letter alphabet, has $k+1$ squares. This is due to the fact that all previous squares cannot reappear later in the word because of the newly introduced letter. □

4 Conclusion

Although the computations done so far show that the actual bound for the one-hole partial words give us at most n distinct squares in any word of length n, the results obtained here using the approach of Fraenkel and Simpson make the bound directly dependable on the size of the alphabet. From our point of view, finding a dependency between the maximum number of squares starting at one position and the length of the word might be a solution. Solving this problem, at least partially, could also give a new perspective to the study of maximum distinct squares within a full word.

Note as well that for arbitrarily large alphabets of size k, we get an upper bound for all words containing h holes and having length n

$$g_{h,k}(n) \leq m_{h,k}(n) + k^{\lfloor \frac{h}{2} \rfloor}$$

This is due to the fact that the leading term is always maximal in $m_{h,k}$, hence adding one to its coefficient we get an upper bound.

References

[1] Berstel, J., Boasson, L. Partial words and a theorem of Fine and Wilf. *Theoretical Computer Science*, 218:135–141, 1999.

[2] Blanchet-Sadri, F. Algorithmic Combinatorics on Partial Words. Chapman & Hall/CRC Press, 2007.

[3] Fraenkel, A.S., Simpson, J. How many squares must a binary sequence contain? *Electronic Journal of Combinatorics*, 2(339 #R2): 1995.

[4] Fraenkel, A.S., Simpson, J. How many squares can a string contain? *Journal of Combinatorial Theory, Series A*, 82:112-120, 1998.

[5] Ilie, L. A simple proof that a word of length n has at most $2n$ distinct squares. *Journal of Combinatorial Theory, Series A*, 112:163-164, 2005.

[6] Ilie, L. A note on the number of squares in a word. *Theoretical Computer Science*, 380:373–376, 2007.

[7] Lothaire, M. Combinatorics on Words. Cambridge University Press, 1997.

[8] Smyth, W.F. Computing Patterns in Strings. Pearson Addison-Wesley, 2003.

Received 16th September 2008