McMaster University

Advanced Optimization Laboratory

Title:

A d-step approach for distinct squares in strings

Authors:

Antoine Deza, Frantisek Franek, and Mei Jiang

AdvOL-Report No. 2011/1

January 2011, Hamilton, Ontario, Canada

A d-step approach for distinct squares in strings. *

Antoine Deza, Frantisek Franek, and Mei Jiang

January 23, 2011

Abstract

We present an approach to the problem of maximum number of distinct squares in a string which underlines the importance of considering as key variables both the length n and $n-d$ where d is the size of the alphabet. We conjecture that a string of length n and containing d distinct symbols has no more than $n-d$ distinct squares, show the critical role played by strings satisfying $n=2d$, and present some properties satisfied by strings of length bounded by a constant times the size of the alphabet.

Keywords: string, distinct squares, primitively rooted distinct squares, d-step approach

1 Introduction

The problem of the number of distinct squares when the types of the squares in a string are counted rather than the occurrences, was first introduced by Fraenkel and Simpson [3] showing that the number of distinct squares in a string of length n is bounded from above by $2n$ and giving a lower bound of $n-o(n)$ asymptotically approaching n from below for primitively rooted squares. Let us remark that a primitively rooted square is a square whose generator is primitive, i.e. not a repetition. Later, Ilie [4] provided a simpler proof of the main lemma of [3] and slightly improved the upper bound to $2n-\Theta(\log n)$ in [5]. It is believed, that the number of distinct squares is bounded by the length of the string.

In this paper we investigate the problem of primitively rooted distinct squares in relationship to the alphabet of the string. Let us denote by $\sigma_d(n)$ the maximum number of primitively rooted distinct squares over all strings of length n containing exactly d distinct symbols. We conjecture that $\sigma_d(n) \leq n-d$, and point to possible avenues for investigating the conjecture.

Similarly as in [2], which was dealing with the maximum number of runs in a string with respect to the string's alphabet, we present some elementary structures of the entries for $\sigma_d(n)$ presented in a so-called (d,n-d) table whose rows are indexed by d and columns are indexed by $n-d$, and point to ways of applying reductions to the problem of distinct squares. A fragment of the table for $d \leq 10$ and $n-d \leq 10$ is shown in Fig. 1.

*Supported in part by grants from the Natural Sciences and Engineering Research Council of Canada for the first 2 authors, and by the Canada Research Chair Programme and Mathematics of Information Technology and Complex Systems grants for the first author, and by Queen Elizabeth II Graduate Scholarship in Science and Technology for the third author.

d \ $n-d$	1	2	3	4	5	6	7	8	9	10	11
1	**1**	***1***	1	1	1	1	1	1	1	1	·
2	1	**2**	***2***	3	3	4	5	6	7	7	·
3	1	2	**3**	***3***	4	4	5	6	7	8	·
4	1	2	3	**4**	***4***	5	5	6	7	8	·
5	1	2	3	4	**5**	***5***	6	6	7	8	·
6	1	2	3	4	5	**6**	***6***	7	7	8	·
7	1	2	3	4	5	6	**7**	***7***	8	8	·
8	1	2	3	4	5	6	7	**8**	***8***	9	·
9	1	2	3	4	5	6	7	8	**9**	***9***	·
10	1	2	3	4	5	6	7	8	9	**10**	·
11	·	·	·	·	·	·	·	·	·	·	·

Figure 1: (d,n-d) table: entries computed for $\sigma_d(n)$ with $1 \leq d \leq 10$ and $1 \leq n-d \leq 10$

Several regularities can be observed in the fragment of the (d,n-d) table: first observe that $\sigma_d(n) \leq n-d$ is satisfied by all known entries. There are several other regularities that can be observed in the table; some are proven analytically in section 2, some are shown to be equivalent with the conjectured upper bound for $\sigma_d(n)$, some are shown to lead to a slightly stronger upper bound – see section 3. In section 4 we investigate the structure of relatively short square-maximal strings on the main diagonal. In section 5, we discuss possible ways to investigate the conjectured upper bound using the methods and insight presented in section 4.

First we introduce the notation used in this paper. $\boldsymbol{S_d(n)}$ denotes the set of strings of length n with exactly d distinct symbols; $\boldsymbol{s(x)}$ denotes the number of primitively rooted distinct squares in a string x; $\boldsymbol{\sigma_d(n)} = \max\{\ s(x) \mid x \in S_d(n)\ \}$. $\mathcal{A}(x)$ denotes the alphabet set of a string x; a ***singleton*** of x refers to a symbol in a string x that occurs exactly once, a ***pair*** refers to a symbol that occurs exactly twice, a ***triple*** refers to a symbol that occurs exactly three times, and in general an ***k-tuple*** (k times).

2 Some basic properties of the (d,n-d) table

The following auxiliary lemma will be used later to investigate the structure of square-maximal strings.

Lemma 1 *Let x be a square-maximal string of length n with exactly d symbols, and let every symbol of x occur at most 2 times. Then every pair in x must be adjacent.*

Proof. Let $x \in S_d(n)$ be square-maximal. Let us assume that x has a non-adjacent pair of C's. Case (i): if the pair does not occur in any square, then we can create a string y by moving the C's to the end. This will not destroy any square of x, but we gain a new square CC, which contradicts the square-maximality of x. Case (ii): if the pair occurs in at least one square, let us move the two C's to the end of the string. For every square $uCvuCv$ of x destroyed by the removal of the C's, we gain a new square $uvuv$: if $uvuv$ already existed in some other part of x, every symbol of uv would have to occur in x at least 3 times, which is not possible. Thus every destroyed square $uCvuCv$ is replaced by a new square $uvuv$, in addition we gain a new square CC. This contradicts the square-maximality of x. $\square$

The next proposition summarizes basic properties of the (d,n-d) table.

Proposition 1 *For any* $2 \leq d \leq n$*:*

(a) $\sigma_d(n) \leq \sigma_d(n+1)$*, i.e. the values are non-decreasing when moving left-to-right along a row.*

(b) $\sigma_d(n) \leq \sigma_{d+1}(n+1)$*, i.e. the values are non-decreasing when moving top-to-bottom along a column.*

(c) $\sigma_d(n) < \sigma_{d+1}(n+2)$*, i.e. the values are strictly increasing when moving left-to-right and top-to-bottom along descending diagonals.*

(d) $\sigma_d(2d) = \sigma_d(n) = \sigma_{d+1}(n+1)$ *for* $n \leq 2d$*, i.e. the values under and on the main diagonal along a column are constant.*

(e) $\sigma_d(n) \geq n-d$ *for* $n \leq 2d$*, i.e. the values under and on the main diagonal are at least as big as conjectured;* $\sigma_d(2d+1) \geq d$ *and* $\sigma_d(2d+2) \geq d+1$*.*

(f) $\sigma_d(2d) - \sigma_{d-1}(2d-1) \leq 1$*, i.e. the difference between the value on the main diagonal and the value immediately above it is no more than 1.*

Proof.

(a) Let $x \in S_d(n)$ be square-maximal. Let y be x appended with a symbol $a \in \mathcal{A}(x)$. Then $y \in S_d(n+1)$, and $\sigma_d(n+1) \geq s(y) \geq s(x) = \sigma_d(n)$.

(b) Let $x \in S_d(n)$ be square-maximal. Let y be x appended with a symbol $a \notin \mathcal{A}(x)$. Then $y \in S_{d+1}(n+1)$, and $\sigma_{d+1}(n+1) \geq s(y) = s(x) = \sigma_d(n)$.

(c) Let $x \in S_d(n)$ be square-maximal, let $a \notin \mathcal{A}(x)$. Define a new string y as x concatenated with aa. Then $y \in S_{d+1}(n+2)$, and $\sigma_{d+1}(n+2) \geq s(y) = s(x)+1 > s(x) = \sigma_d(n)$.

(d) Let $n \leq 2d$ and let $x \in S_{d+1}(n+1)$ be square-maximal. Since $2(d+1) \geq n+2 > n+1$, x has a singleton. Let y be x with the singleton removed. Then $y \in S_d(n)$ and $s(y) \geq s(x)$ as no square can be destroyed while some squares can be created. Thus, $\sigma_d(n) \geq s(y) \geq s(x) = \sigma_{d+1}(n+1)$. By (b), $\sigma_d(n) \leq \sigma_{d+1}(n+1)$, so $\sigma_d(n) = \sigma_{d+1}(n+1)$ for $n \leq 2d$.

(e) Let $n \leq 2d$ and consider the string $x = aabbcc\ldots$ consisting of $n-d$ adjacent pairs. Then $x \in S_{n-d}(2n-2d)$ and $s(x) = n-d$. By (d), $\sigma_d(n) = \sigma_{n-d}(2n-2d) \geq s(x) = n-d$. Let consider the strings $y = aaabbcc\ldots$ consisting of $d-1$ adjacent pairs except for the first 3 entries being aaa, and $z = aababaccdd\ldots$ consisting of $d-2$ adjacent pairs except for the first 6 entries being $aababa$. We have $\sigma_d(2d+1) \geq s(y) = d$ and $\sigma_d(2d+2) \geq s(z) = d+1$.

(f) Let $x \in S_d(2d)$ be square-maximal. Case (i): if x has a singleton, let y be x with the singleton removed, then $y \in S_{d-1}(2d-1)$ and $s(y) \geq s(x)$. It follows that $\sigma_d(2d) = s(x) \leq s(y) \leq \sigma_{d-1}(2d-1)$, and since $\sigma_d(2d) \geq \sigma_{d-1}(2d-1)$ by (b), therefore we get $\sigma_d(2d) = \sigma_{d-1}(2d-1)$. Case (ii): if x does not have a singleton, then x consists of pairs, and by Lemma 1, x consists of adjacent pairs, and thus $\sigma_d(2d) = s(x) = d$. Consider the string $z = aaabbcc\ldots$ consisting of $d-2$ adjacent pairs except for the first 3 entries being aaa. We have $\sigma_{d-1}(2d-1) \geq s(z) = d-1 = \sigma_d(2d) - 1$, i.e., $\sigma_d(2d) - \sigma_{d-1}(2d-1) \leq 1$. □

3 Main results

This sections contains several propositions that are equivalent with the conjectured upper bound for $\sigma_d(n)$. We also present conditions that lead to a slightly stronger upper bound in Theorems 3 and 4. It can be observed in the (d,n-d) table, that the known values on the main diagonal are identities, i.e. $\sigma_d(2d) = d$ – which is equivalent to $\sigma_d(2d) \leq d$ by Proposition 1(e). The next theorem shows that, indeed, this observation is equivalent with the conjectured bound. In essence, the theorem shows that if the upper bound is violated, then there must be a violation on the main diagonal.

Theorem 1 *The conjectured upper bound $\sigma_d(n) \leq n - d$ holding true for all strings is equivalent with the statement:* $\sigma_d(2d) \leq d$ for every $d \geq 2$.

Proof. Let $n \geq d \geq 2$, $\sigma_d(n) \leq n - d$ clearly implies that $\sigma_d(2d) \leq d$; that is, by Proposition 1(e), $\sigma_d(2d) = d$. To prove the other direction, we consider case (i) $2d > n$: by Proposition 1(d) we have $\sigma_d(n) = \sigma_{n-d}(2n-2d) \leq n-d$, and case (ii) $n > 2d$: by Proposition 1(b) we have $\sigma_d(n) \leq \sigma_{n-d}(2n-2d) \leq n-d$. □

Another observation of the (d,n-d) table given in Figure 1 is that the value on the main diagonal and the value of its right neighbour are identical. Theorem 2 shows that the inequality is equivalent with the conjectured upper bound, while the equality gives rise to a slightly stronger upper bound given in Theorem 4.

Theorem 2 *The conjectured upper bound $\sigma_d(n) \leq n - d$ holding true for all strings is equivalent with the statement:* $\sigma_d(2d+1) - \sigma_d(2d) \leq 1$ for every $d \geq 2$.

Proof. The statement follows from the conjectured upper bound is clear. Let us, thus prove the opposite direction. We shall prove by contradiction that $\sigma_d(2d) \leq d$ for $d \geq 2$. Let $d \geq 2$ be the least such that $\sigma_d(2d) > d$. From the computed values of the $(d, n-d)$ table it follows that $d > 10$. Let $x \in S_d(2d)$ be square-maximal. If x does not have a singleton, then $n = 2d$ and x consists of pairs, and thus by Lemma 1, x consists of adjacent pairs and $\sigma_d(2d) = d$, a contradiction. Thus, x must have a singleton. Let y be x with the a singleton removed. Then $y \in S_{d-1}(2d-1)$ and $s(y) \geq s(x)$. Thus, $\sigma_{d-1}(2d-1) \geq s(y) \geq s(x) = \sigma_d(2d)$. Moreover,

$\sigma_{d-1}(2d-1) \le \sigma_{d-1}(2d-2)+1 \le d-1+1 = d$. Thus, $d \ge \sigma_{d-1}(2d-1) = \sigma_d(2d) > d$, a contradiction. Therefore, $\sigma_d(2d) \le d$ for every $d \ge 2$ and the conjectured upper bound follows by applying Theorem 1. □

Another observation of the (d,n-d) table given in Figure 1 is that not only $\sigma_d(2d)$ is bounded by d, but also it is true for $\sigma_d(2d+1)$. Theorem 3 shows that this property implies a slightly stronger upper bound.

Theorem 3 *If $\sigma_d(2d+1) \le d$ for every $d \ge 2$, then $\sigma_d(n) \le n-d-1$ for $n > 2d \ge 4$ and $\sigma_d(n) = n-d$ for $n \le 2d$.*

Proof. We have $d \le \sigma_d(2d) \le \sigma_d(2d+1) \le d$ and so $\sigma_d(2d) = \sigma_d(2d+1) = d$. It implies that $\sigma_d(n) = n-d$ for $n \le 2d$. For $n > 2d$ we have, by Proposition 1(b), $\sigma_d(n) \le \sigma_{n-d-1}(2n-2d-1) \le n-d-1$. □

Theorem 4 *If $\sigma_d(2d) = \sigma_d(2d+1)$ for every $d \ge 2$, then $\sigma_d(n) \le n-d-1$ for $n > 2d \ge 4$ and $\sigma_d(n) = n-d$ for $n \le 2d$.*

Proof. The results follow from Theorem 3 and the fact that $\sigma_d(2d) = \sigma_d(2d+1) = d$ for every $d \ge 2$. To show that $\sigma_d(2d) = \sigma_d(2d+1) = d$ for every $d \ge 2$, let us argue by contradiction. Let d be the smallest such that $\sigma_d(2d) = \sigma_d(2d+1) > d$. From the values in the (d,n-d) table calculated so far, we know that $d > 10$. Thus $d-1 = \sigma_{d-1}(2d-2) = \sigma_{d-1}(2d-1)$. However, by Proposition 1(f), $\sigma_{d-1}(2d-1)+1 \ge \sigma_d(2d)$. It follows that $d-1 \ge \sigma_d(2d)-1$. i.e. $d \ge \sigma_d(2d)$, a contradiction. □

4 Structure of relatively short square-maximal strings

In this section we investigate square-maximal strings that are short relative to the size of their alphabets. The main goal of this investigation is to either find a counterexample on the main diagonal if there is one, or to show that there are no counterexamples on the main diagonal, as this would prove the conjectured upper bound for all strings. We show that a square-maximal string from the main diagonal either complies with the conjectured upper bound or has to have many singletons based on the facts that such string (a) cannot contain pairs, see Lemma 4, and (b) if it contains a triple, it is must be a very special triple, implying the existence of a symbol occurring at least 6 times, see Lemma 8. We hope that it might be possible to show that counterexamples on the main diagonal do not exist by showing that their structure would be impossible. We discuss this in Conclusion.

Lemma 2 shows the structure of the square-maximal strings on the main diagonal if they are in compliance with the conjectured upper bond and they are identical with the value of its right neighbour.

Lemma 2 *If $\sigma_d(2d) = \sigma_d(2d+1)$ for every $d \ge 2$, then for any $d \ge 2$, $x \in S_d(2d)$ square-maximal, x is up to relabeling of the alphabet, unique and equal to $x = (aabbcc\dots)$.*

Proof. If x contains only pairs, by Lemma 1 all these pairs have to be adjacent. If x did not consist only of pairs, then it would have to have a singleton. Let y be a string obtained from x by removing a singleton. $y \in S_{d-1}(2d-1)$ and $s(y) \geq s(x)$. Thus $d-1 = \sigma_{d-1}(2d-2) = \sigma_{d-1}(2d-1) \geq s(y) \geq s(x) = \sigma_d(2d) = d$ which is contradiction. Therefore x contains only pairs and is up to relabeling, unique and equal to $x = (aabbcc\ldots)$. □

Auxiliary Lemma 3 will be used to estimate the number of squares that span from one part of a string to the other part and relies on the result of Fraenkel and Simpson [3].

Lemma 3 *Consider non-empty strings w, u, and v. The number of distinct squares of the string wuv that start in w and end in v is at most $|w|+|v|$ where $|w|$, respectively $|v|$, denotes the length of w, respectively v.*

Proof. We discuss two cases: Case (i) $|w| \leq |v|$: we count the rightmost occurrences of squares. By Fraenkel-Simpson [3], there are at most two such squares starting at the same position. Thus, there are at most $2|w|$ squares that start in w, and $2|w| \leq |w| + |v|$. Case (ii) $|w| > |v|$: let $\overline{x}$ denote the reversal of the string x. By the previous argument, there are at most $2|\overline{v}|$ squares of the string $\overline{wuv} = \overline{v}\ \overline{u}\ \overline{w}$ starting in $\overline{v}$. It follows that there are at most $2|v|$ squares of wuv that end in v and $2|v| < |w| + |v|$. □

Lemma 4 shows that the square-maximal strings in first unknown position on the main diagonal either comply with the conjectured upper bound or cannot contain a pair.

Lemma 4 *Let $\sigma_{d'}(2d') \leq d'$ where $d' < d$. Let $x \in S_d(2d)$ be square-maximal. Then either $s(x) = \sigma_d(2d) = d$ or x does not contain a pair.*

Proof. Let assume that $s(x) = \sigma_d(2d) > d$ and x contains a pair of C's at positions i_0 and i_1, so $x[i_0] = x[i_1] = C$. If the pair occurs in at most 1 square, then we can replace the first C with a new symbol $\hat{C} \notin \mathcal{A}(x)$. Let y be x with $x[i_0]$ replaced by $\hat{C}$. Then $y \in S_{d+1}(2d)$ and $\sigma_{d+1}(2d) \geq s(y) = s(x) - 1 = \sigma_d(2d) - 1$. Since $2d - (d+1) < d$, we get $2d - (d+1) \geq \sigma_{d+1}(2d) \geq \sigma_d(2d) - 1$, i.e. $d - 1 \geq s(x) - 1$, and so $d \geq s(x)$, a contradiction. Therefore, the pair must occur in at least two squares, in fact in a non-trivial run $x = \cdots uvCwuvCwu \cdots$, where $|u| \geq 1$. Let us form a new string y by removing all the symbols between the C's: $y = \cdots uvCCwu \cdots$. By doing this, we may have destroyed $|u|+1$ squares – $uvCwuvCw$ and its $|u|$ rotations. The type of any square of u is preserved, as y has u as a substring. The same is true for w, v, wu, and uv. Thus, we may have destroyed the squares of wuv that start in w and end in v. By Lemma 3, we may have destroyed at most $|w| + |v|$ squares. So, altogether, we may have destroyed at most $|w| + |u| + |v| + 1$ squares, but we created a new one: CC. Thus $s(y) \geq s(x) - (|w| + |u| + |v|)$. Clearly, $\mathcal{A}(y) = \mathcal{A}(x)$, and so $y \in S_d(2d-k)$ where $k = |w| + |u| + |v|$. By the assumption of this lemma as $2d - k - d = d - k < d$, we have $d - k \geq \sigma_d(2d-k) \geq s(y) \geq s(x) - k$, and thus $d \geq s(x)$, a contradiction. □

Lemmas 5 and 6 use the same scenario investigating the square-maximal strings in the first unknown position on the main diagonal and showing that they either comply with the conjectured upper bound or may contain only very specific triples.

Lemma 5 *Let $\sigma_{d'}(2d') \leq d'$ where $d' < d$. Let $x \in S_d(2d)$ be square-maximal. Then either $s(x) = \sigma_d(2d) = d$ or if x contains a triple, then the triple has to occur in two distinct runs.*

Proof. Let assume that $s(x) = \sigma_d(2d) > d$. Let $x[i_0] = x[i_1] = x[i_2] = C$ be a triple in x. We first show all three symbols occur in some runs. Assume that $x[i_0]$ does not occur in any run. Let $\hat{C}$ be a symbol $\notin \mathcal{A}(x)$. Let y be x with $x[i_0]$ replaced by $\hat{C}$. Then $y \in S_{d+1}(2d)$ and $\sigma_{d+1}(2d) \geq s(y) = s(x) = \sigma_d(2d)$. Since $2d - (d+1) < d$, we get $2d-(d+1) \geq \sigma_{d+1}(2d) \geq \sigma_d(2d)$, i.e. $d-1 \geq \sigma_d(2d)$, a contradiction. For $x[i_2]$ not occurring in any run, the proof is the same. If $x[i_1]$ does not occur in any run, then none of the elements of the triple occur in any run. Then we can remove $x[i_1]$ forming a string $y \in S_d(2d-1)$ such that $d-1 \geq \sigma_d(2d-1) \geq s(y) \geq s(x) = \sigma_d(2d)$, a contradiction. We then show the three symbols cannot occur in the same run. Assume they do occur in the run $uvCwuvCwuvCwu$. We can proceed as in the proof of Lemma 4 and remove wuv between the first and second C. □

Lemma 6 *Let $\sigma_{d'}(2d') \leq d'$ where $d' < d$. Let $x \in S_d(2d)$ be square-maximal. Then either $s(x) = \sigma_d(2d) = d$, or if x has a triple $x[i_0] = x[i_1] = x[i_2] = C$ occurring in two distinct runs $u_1v_1x[i_0]w_1u_1v_1x[i_1]w_1u_1 = u_1v_1Cw_1u_1v_1Cw_1u_1$ and $u_2v_2x[i_1]w_2u_2v_2x[i_2]w_2u_2 = u_2v_2Cw_2u_2v_2Cw_2u_2$, then $|u_1| \geq 1$ and $|u_2| \geq 1$ and either u_2v_2 is not a suffix of u_1v_1 or w_1u_1 is not a prefix of w_2u_2.*

Proof. Let us assume that $s(x) = \sigma_d(2d) > d$. If $|u_1| = 0$, then $x[i_0]$ occurs in a single square $v_1Cw_1v_1Cw_1$. Let $\hat{C}$ be a symbol $\notin \mathcal{A}(x)$ and let y be x with $x[i_0]$ replaced by $\hat{C}$. Then $y \in S_{d+1}(2d)$ and $\sigma_{d+1}(2d) \geq s(y) = s(x) - 1 = \sigma_d(2d) - 1$. Since $2d - (d+1) < d$, we get $2d-(d+1) \geq \sigma_{d+1}(2d) \geq \sigma_d(2d)-1$, i.e. $d-1 \geq \sigma_d(2d)-1$, and so $d \geq \sigma_d(2d)$, a contradiction. It follows that $|u_1| \geq 1$. For $|u_2| = 0$, the proof is the same. Thus, $|u_1| \geq 1$ and $|u_2| \geq 1$. Let us assume that both u_2v_2 is a suffix of u_1v_1 and w_1u_1 a prefix of w_2u_2. Let us form a new string y from x by removing $w_1u_1v_1$ between $x[i_0]$ and $x[i_1]$ and removing $w_2u_2v_2$ between $x[i_1]$ and $x[i_2]$, that is $y = x[1..i_0]x[i_1]x[i_2..2d] = x[1..i_0-1]CCCx[i_2+1..2d]$. It follows that $y \in S_d(2d-k)$ where $k = |w_1| + |u_1| + |v_1| + |w_2| + |u_2| + |v_2|$. How many squares we might have destroyed? We might have destroyed $|u_1|+1$ squares of $u_1v_1Cw_1u_1v_1Cw_1u_1$ and $|u_2|+1$ squares of $u_2v_2Cw_2u_2v_2Cw_2u_2$. From $w_1u_1v_1$, u_1v_1 has been preserved, w_1u_1 is a prefix of w_2u_2 that was preserved, so the only squares we might have destroyed are the ones starting in w_1 and ending in v_1, and by Lemma 3 there are at most $|w_1| + |v_1|$ of them. Similarly for $w_2u_2v_2$. Thus we might have destroyed at most $|w_1| + |u_1| + |v_1| + |w_2| + |u_2| + |v_2| + 2 = k+2$ squares, and we gained one (CCC). It follows that $s(y) \geq s(x) - k - 1$. Replace the first C in y by a new symbol $\hat{C} \notin \mathcal{A}(x)$ to form a string z. Then $z \in S_{d+1}(2d-k)$ and $s(z) = s(y)$. Thus $\sigma_{d+1}(2d-k) \geq s(z) = s(y) \geq s(x) - k - 1 = \sigma_d(2d) - k - 1$. Since $2d - k - d - 1 = 2d - |w_1| - |u_1| - |v_1| - |w_2| - |u_2| - |v_2| - d - 1 < d$, we have $2d-k-d-1 \geq \sigma_{d+1}(2d-k) \geq s(x)-k-1$, so $2d-k-d-1 \geq s(x)-k-1$ and so $d \geq s(x)$, a contradiction. It follows that either u_2v_2 is not a suffix of u_1v_1, in which case u_1v_1 is a suffix of u_2v_2, or w_1u_1 is not a prefix of w_2u_2, in which case w_2u_2 is a prefix of w_1u_1. □

Lemma 7 shows that the square-maximal strings cannot contain parallel k-tuples. A k-tuple of C's occurring at positions $\{i_1, \cdots i_k\}$ and a k-tuple of D's occurring at positions $\{j_1, \cdots j_k\}$ are *parallel* if $i_1 < j_1 < i_2 < j_2 < \cdots < i_k < j_k$.

Lemma 7 *Let $x \in S_d(2d)$ be square-maximal. Then x cannot contain two parallel k-tuples for any $k \geq 2$.*

Proof. Let us assume that x contains two parallel k-tuples of C's and D's. Let us move all D's to the end of the string x, forming a new string $y \in S_d(2d)$. Any primitively rooted square that contains m of the D's must also contain at least m of the C's. If we remove the D's from the square, we create a new square. Since it contains the C's and since the original square was primitively rooted, the new square also must be primitively rooted. For illustration: $[uCvDw][uCvDw]$ will become $[uCvw][uCvw]$. Moving the D's to the end creates a new square DD and so $s(y) > s(x)$, a contradiction with the square-maximality of x. □

Lemma 8 utilizes the previous lemmas and shows that any square-maximal string in the first unknown position on the main diagonal either complies with the conjectured upper bound, or if if it contains a triple, it must be a very specific one giving rise to a symbol that must occur at least 6 times. Thus, each triple occurring must be balanced by an existence of a unique set of 5 occurrences of a certain symbol. Though the symbol may not be unique to a particular triple, the set of occurrences are mutually disjoint. Thus, every triple with its assigned set of 5 occurrences is balanced by an existence of at least 4 singletons unique to the triple and its assigned set.

Lemma 8 *Let $\sigma_{d'}(2d') \leq d'$ where $d' < d$. Let $x \in S_d(2d)$ be square-maximal. Then either $s(x) = \sigma_d(2d) = d$ or x has at least $\lceil \frac{2d}{3} \rceil$ singletons.*

Proof. Let us assume that $s(x) = \sigma_d(2d) > d$. From Lemma 4 it follows, that x does not have any pair. From Lemmas 5 and 6, any triple $x[i_0] = x[i_1] = x[i_2] = C$ of x must be special, i.e. it must satisfy

1. $x[i_0]$ and $x[i_1]$ occur in a run $r_1 = u_1v_1Cw_1u_1v_1Cw_1u_1$, where $|u_1| \geq 1$,
2. $x[i_1]$ and $x[i_2]$ occur in a run $r_2 = u_2v_2Cw_2u_2v_2Cw_2u_2$, where $|u_2| \geq 1$, and where $i_1 - i_0 \neq i_2 - i_1$ as otherwise the two runs would merge into a single one,
3. either u_1v_1 is a proper suffix of u_2v_2, or w_2u_2 is a proper prefix of w_1u_1.

Let us discuss the case when u_1v_1 is a proper suffix of u_2v_2; the case of w_2u_2 being a proper prefix of w_1u_1 is the same just argued from the opposite direction. Let the run $r_1 = u_1v_1Cw_1u_1v_1Cw_1u_1$ start at position t of x. Consider $a = x[t]$. If there is no occurrence of a in $x[t+1..i_0-1]$, then we can replace all occurrences of a in $x[1..i_0-1]$ with a new symbol, forming a string y, while destroying a single square $u_1v_1Cw_1u_1v_1Cw_1$ of x. Thus $y \in S_{d+1}(2d)$, $2d - d - 1 \geq \sigma_{d+1}(2d) \geq s(y) = s(x) - 1 = \sigma_d(2d) - 1$, so $d \geq \sigma_d(2d)$, a contradiction. Thus a occurs at least twice in $x[t..i_0-1] = u_1v_1$. Since u_1v_1 is a suffix of u_2v_2, a occurs at least 4 more times – twice in each occurrence of u_2v_2. Thus, $x[t]$ occurs in x at least six times, the last occurrence before the last C. We assign to the triple the sequence of positions of the 5 first occurrences of a after the position t and denote it by $As(C) = \langle j_0, j_1, j_2, j_3, j_4 \rangle$, where $t < j_0 < j_1 < j_2 < j_3 < j_4 < i_2$ and $j_0 < i_0$ and t is the start of the run r_1 and $x[t] = x[j_0] = x[j_1] = x[j_2] = x[j_3] = x[j_4]$. Of course, if the short appendix used was w_2u_2, then $As(C) = \langle j_0, j_1, j_2, j_3, j_4 \rangle$, where $i_0 < j_4 < j_3 < j_2 < j_1 < j_0 < t$ and $i_2 < j_0$ and t is the end of the run r_2 and $x[t] = x[j_0] = x[j_1] = x[j_2] = x[j_3] = x[j_4]$. Below, we will show that such assignments are mutually disjoint, i.e. if C's and D's are different triples, then $As(C) \cap As(D) = \emptyset$.

Now we can estimate the number of singletons in x. Let m_0 be the number of triples in x. Let m_1 be the number of multiply occurring symbols that are not assigned to triples – since there are no pairs, it follows that such symbols occur at least 4 times. Finally, let m_2 be the number of singletons in x. The following 2 inequalities must hold: $2d \geq 8m_0 + 4m_1 + m_2$ and $d \leq 2m_0 + m_1 + m_2$ which clearly yields $3m_2 \geq 2d$ and so $m_2 \geq \lceil \frac{2d}{3} \rceil$.

A proof of the claim that the assignments are mutually disjoint: Let $As(C) = \langle j_0, j_1, j_2, j_3, j_4 \rangle$ and let $As(D) = \langle k_0, k_1, k_2, k_3, k_4 \rangle$. If $x[j_0] \neq x[k_0]$, then $As(C) \cap As(D) = \emptyset$. Bellow, we discuss the case when $x[j_0] = x[k_0] = a$.

In Lemma 6 it is shown that a triple of C's can exist in x only if it occurs in two distinct non-trivial runs $u_1v_1Cw_1u_1v_1Cw_1u_1$ and $u_2v_2Cw_2u_2v_2Cw_2u_2$. We refer to u_1v_1 and w_2u_2 as the appendices, and we say that u_1v_1 is a short appendix if u_1v_1 is a proper suffix of u_2v_2, similarly we say that w_2u_2 is a short appendix if it is a proper prefix of w_1u_1. Thus, Lemma 6 also stipulates that at least one of the appendices must be short.

Let us consider two different triples, one of C's and one of D's and let us assume that the first C precedes the first D. We must discuss all the possible configurations of the two triples. For better readability, we will denote by C_1 the first occurrence of C, by C_2 the second occurrence of C etc. Similarly for D's.

The C's occur in two non-trivial runs $r_1 = u_1v_1C_1w_1u_1v_1C_2w_1u_1$ and $r_2 = u_2v_2C_2w_2u_2v_2C_3w_2u_2$, while the D's occur in two non-trivial runs $r_3 = u_3v_3D_1w_3u_3v_3D_2w_3u_3$ and $r_4 = u_4v_4D_2w_4u_4v_4D_3w_4u_4$.

1. C_3 occurs before D_1, i.e. the triples do not interleave (schematically $C_1\ C_2\ C_3\ D_1\ D_2\ D_3$).
   (a) First we consider the case when the appendix determining $As(C)$ and the appendix determining $As(D)$ are on the opposite sides.
   Thus, the short appendix determining $As(C)$ is on the left and the short appendix determining $As(D)$ is on the right. Then we are guarantied the following pattern of occurrences of a in x (for the C's, the a's are shown in bold, for the D's, the a's are shown underscored): $x = \cdots \boldsymbol{a}\ \boldsymbol{a}\ C_1\ \boldsymbol{a}\ \boldsymbol{a}\ C_2\ \boldsymbol{a}\ \boldsymbol{a}\ C_3\ D_1\ \underline{a}\ \underline{a}\ D_2\ \underline{a}\ \underline{a}\ D_3\ \underline{a}\ \underline{a} \cdots$, so $x[j_4]$ occurs before C_3, while the $x[k_4]$ occurs after D_1. Therefore $j_4 < k_4$ and so $As(C) \cap As(D) = \emptyset$.
   (b) Next we consider the case when the appendix determining $As(C)$ and the appendix determining $As(D)$ are facing each other.
   Thus, for the C's we are using the right appendix, for the D's the left appendix. Then we are guarantied the following pattern of occurrences of a in x (for the C's, the a's are shown in bold, for the D's, the a's are shown underscored):
   $x = \cdots C_1\ \boldsymbol{a}\ \boldsymbol{a}\ C_2\ \boldsymbol{a}\ \boldsymbol{a}\ C_3\ \underline{\boldsymbol{a}}\ \underline{\boldsymbol{a}}\ D_1\ \underline{a}\ \underline{a}\ D_2\ \underline{a}\ \underline{a}\ D_3 \cdots$, and thus $x[j_0]$ occurs at or to the left of $\boldsymbol{a}$ (shown in bold), while $x[k_0]$ occurs at or to the right of $\underline{a}$ (shown underscored). It is possible that two a's between C_3 and D_1 are the same. However, since we do not take the first occurrence of a for the assignments, $As(C) \cap As(D) = \emptyset$.

(c) Here we consider the case when the appendix determining $As(C)$ and the appendix determining $As(D)$ are on the same side.
Without loss of generality, we can assume that both appendices used are on the left. Then we are guarantied the following pattern of occurrences of a in x (for the C's, the a's are shown in bold, for the D's, the a's are shown underscored):
$x = \cdots \boldsymbol{a}\ \boldsymbol{a}\ C_1\ \boldsymbol{a}\ \boldsymbol{a}\ C_2\ \boldsymbol{a}\ \boldsymbol{a}\ C_3\ \underline{a}\ \underline{a}\ D_1\ \underline{a}\ \underline{a}\ D_2\ \underline{a}\ \underline{a}\ D_3 \cdots$.Why cannot the first two $\underline{a}$'s be the same as the last two $\boldsymbol{a}$'s? If it were the case, then C would be in the appendix for the D's, i.e. a part of the run r_3 and hence repeat later. So, again $As(C) \cap As(D) = \emptyset$.

2. Case $x = \cdots C_1\ D_1\ D_2\ C_2 \cdots$ is not possible.
If either D_1 or D_2 occurred in u_1v_1, then there would be a D preceding C_1. Thus both D_1 and D_2 occur in w_1, but then D occurs at least 4 times, a contradiction.
3. Case $x = \cdots C_1\ D_1\ C_2\ D_2\ D_3\ C_3 \cdots$ is not possible.
As in the previous case, D_1 must occur in w_1 and D_2 together with D_3 must occur in v_2, hence D must occur at least 4 times, a contradiction.
4. Case $x = \cdots C_1\ D_1\ C_2\ D_2\ C_3\ D_3 \cdots$.
This case is not possible by Lemma 7 as the triples of C's and D's are parallel.
5. Case $x = \cdots C_1\ C_2\ D_1\ C_3\ D_2\ D_3 \cdots$.
We denote by $w_2^{(1)}$ the first occurrence of w_2 in x, by $w_2^{(2)}$ the second occurrence of w_2 in x, etc.
If D_1 occurred in $(u_2v_2)^{(2)}$, there would be a D preceding C_2. Hence D_1 must occur in $w_2^{(1)}$ and hence D_2 occurs in $w_2^{(2)}$. Since the distance between C_2 and C_3 is the period of r_2, and the distance between D_1 and D_2 is the period of r_3, and the distances are equal, it follows that $r_2 = r_3 = u_2v_2C_2w'_2D_1w''_2u_2v_2C_3w'_2D_2w''_2u_2$ (note that $u_3 = u_2$ and $v_3 = v_2Cw'_2$ and $w_3 = w''_2$.)
Schematically:

$$\begin{array}{lll} r_1: & u_1v_1C_1w_1u_1v_1C_2w_1u_1 & \\ r_2 = r_3: & \quad u_2v_2C_2w'_2D_1w''_2u_2v_2C_3w'_2D_2w''_2u_2 & \\ r_4: & \qquad\qquad u_4v_4D_2w_4u_4v_4D_3w_4u_4 & \end{array}$$

Now consider the two runs r_1 and r_2. Since D_1 cannot occur in $(w_1u_1)^{(2)}$, it follows that the w_1u_1 is a prefix of w'_2 and hence of $w'_2D_1w''_2u_2$, and so the appendix $w'_2D_2w''_2u_2$ is long and by Lemma 6, u_1v_1 must be a short appendix and is used to determine $As(C)$.
Now consider the two runs r_3 and r_4. Since C_3 cannot occur in $(u_4v_4)^{(1)}$, u_4v_4 is a suffix of w'_2 and hence of $u_2v_2C_3w'_2$, and so the appendix $u_2v_2C_2w'_2$ is long. By Lemma 6, w_4u_4 must be a short appendix and is used to determine $As(D)$.

(a) Let a occur twice in u_1 and in twice in u_4 (*the dots indicate the occurrences*).

$$\begin{array}{lll} r_1: & \ddot{u}_1v_1C_1w_1\ddot{u}_1v_1C_2w_1\ddot{u}_1 & \\ r_2 = r_3: & \quad u_2v_2C_2w'_2D_1w''_2u_2v_2C_3w'_2D_2w''_2u_2 & \\ r_4: & \qquad\qquad \underset{..}{u_4}v_4D_2w_4\underset{..}{u_4}v_4D_3w_4\underset{..}{u_4} & \end{array}$$

Then a occurs twice in each occurrence of u_1 and hence $x[j_4]$ occurs in or before $u_1^{(3)}$. Similarly, a occurs twice in each occurrence of u_4 and hence $x[k_4]$ occurs in or after $u_4^{(1)}$. Thus $As(C) \cap As(D) = \emptyset$.

(b) Let a occur only once in u_1 and twice in u_4.

$$\begin{array}{ll} r_1: & \dot{u}_1\dot{v}_1C_1w_1\dot{u}_1\dot{v}_1C_2w_1\dot{u}_1 \\ r_2 = r_3: & \quad u_2v_2C_2{w'}_2D_1{w''}_2\dot{u}_2\dot{v}_2C_3{w'}_2D_2{w''}_2u_2 \\ r_4: & \qquad \underset{\cdot\cdot}{u_4}v_4D_2w_4\underset{\cdot\cdot}{u_4}v_4D_3w_4\underset{\cdot\cdot}{u_4} \end{array}$$

Then a must occur in v_1. Since u_1v_1 is a suffix of u_2v_2 and since w_1u_1 is a prefix of w_2', we have 7 occurrences of a from the left and 6 occurrences of a from the right, so again $As(C) \cap As(D) = \emptyset$.

(c) Let a occur twice in u_1 and only once in u_4.
This is symmetric to the previous case, we will have 6 occurrences of a from the left, and 7 occurrences of a from the right.

(d) Let a occur in u_1 only once and in u_4 also only once.

$$\begin{array}{ll} r_1: & \dot{u}_1\dot{v}_1C_1w_1\dot{u}_1\dot{v}_1C_2w_1\dot{u}_1 \\ r_2 = r_3: & \quad u_2v_2C_2\underset{\cdot}{{w'}_2}D_1\underset{\cdot}{{w''}_2}\underset{\cdot}{\dot{u}_2}\dot{v}_2C_3\dot{w}'_2D_2{w''}_2u_2 \\ r_4: & \qquad \underset{\cdot}{u_4}v_4D_2\underset{\cdot}{w_4}\underset{\cdot}{u_4}v_4D_3\underset{\cdot}{w_4}\underset{\cdot}{u_4} \end{array}$$

From the left there are 8 occurrences of a: a must occur in v_1 and since u_1v_1 is a suffix of u_2v_2, it must occur twice in $(u_2v_2)^{(2)}$, and since u_1 is a substring of w_2', a must occur in all occurrences of w_2'. Similarly, there are 8 occurrences of a from the right. Even though it is possible the the last four occurrences from the left and the last four occurrences from the right are the same, the first 6 occurrences from the left and 6 occurrences from the right are disjoint, and so $As(C) \cap As(D) = \emptyset$.

□

Theorem 5 stresses the fact that the first position on the main diagonal violating the conjectured upper bound implies an existence of a counterexample higher up. Similarly as Theorems 1 and 2, this is yet another reformulation of the conjectured upper bound.

Theorem 5 *The conjectured upper bound $\sigma_d(n) \leq n - d$ holding true for all strings is equivalent with the statement:* $\sigma_d(4d) \leq 3d$ for every $d \geq 2$.

Proof. The statement clearly follows from the conjectured upper bound. We shall prove the opposite direction by contradiction. Let us assume that the conjectured upper bound does not hold. By Theorem 1, it follows that there is a counterexample x on the main diagonal, i.e. a square-maximal $x \in S_d(2d)$ with $s(x) = \sigma_d(2d) > d$. Let us consider the first column d of the table in which the counterexample occurs, from the table as computed so far, we know that $d > 10$. By Lemma 8, x has at least $\lceil \frac{2d}{3} \rceil$ singletons. If we remove $\lceil \frac{2d}{3} \rceil$ singletons from x, we get a string $y \in S_{d'}(n')$ such that $s(y) \geq s(x) > d$ where $d' = d - \lceil \frac{2d}{3} \rceil$ and $n' = 2d - \lceil \frac{2d}{3} \rceil$. Moreover, $4d' = 4(d - \lceil \frac{2d}{3} \rceil) = 4d - 4 \cdot \lceil \frac{2d}{3} \rceil = 4d - 2d - \lceil \frac{2d}{3} \rceil = 2d - \lceil \frac{2d}{3} \rceil = n'$, thus $n' = 4d'$. So we have $\sigma_{d'}(4d') \geq s(y) \geq s(x) = \sigma_d(2d) > d$ and since $3d' = 4d' - d' =$

$n' - d' = (2d - \lceil \frac{2d}{3} \rceil) - (d - \lceil \frac{2d}{3} \rceil) = d$, $\sigma_{d'}(4d') > 3d'$. Thus, we have a counterexample from $S_{d'}(4d')$. □

5 Conclusions

The methods used in section 4 illustrate two possible approaches to investigate the conjectured upper bound for all strings. One is to show that the first counterexample on the main diagonal cannot have a pair, a triple, a quadruple, ... or an k-tuple, i.e. it cannot exist. This approach is represented by Lemma 4. The other approach is to show that if the first counterexample on the main diagonal contains a k-tuple, then it must contain a symbol with a frequency $> k$. This also leads to the conclusion that a counterexample cannot exist. This approach is represented by the proof of Lemma 8. Thus, Lemmas 4 and 8 illustrate the usefulness of investigating the more orderly world of the strings on the main diagonal.

Let us just remark that our approach was inspired by a similar (d,n-d) table used for investigating the Hirsch bound for the diameter of bounded polytopes. The associated Hirsch (d,n-d) table exhibits similar regularities as the (d,n-d) table considered in this paper. The Conjecture of Hirsch was recently disproved by Santos [7] by exhibiting a violation on the main diagonal with $d = 43$ which was further improved to $d = 20$, see [6]. Similarly, we hope that the structure of square-maximal strings is richer for $n = 2d$ and therefore this could be the focus of investigation for tackling the conjectured upper bound. For instance, while for known values there is only essentially a single square-maximal string on the main diagonal and it has a well-described structure, the further up from the diagonal, the more irregular and unpredictable the set of square-maximal strings and their structures are.

An analogue of Theorem 5 for the maximal number of runs given in [1] shows that the conjectured upper bound of $n - d$ for the number of runs holding true for all strings is equivalent with the upper bound of $8d$ for strings in $S_d(9d)$ for every $d \geq 2$.

References

[1] A. Baker, A. Deza, and F. Franek, *On the structure of relatively short run-maximal strings*, AdvOL Technical Report 2011/02, Department of Computing and Software, McMaster University, Hamilton, Ontario, Canada.

[2] A. Deza and F. Franek, *A d-step analogue for runs on strings*, AdvOL Technical Report 2010/02, Department of Computing and Software, McMaster University, Hamilton, Ontario, Canada.

[3] A. S. Fraenkel and J. Simpson, *How Many Squares Can a String Contain?*, Journal of Combinatorial Theory Series A, 82, 1 (1998), 112-120.

[4] L. Ilie, *A simple proof that a word of length n has at most 2n distinct squares*, Journal of Combinatorial Theory Series A, 112, 1 (2005) 163-164.

[5] L. Ilie, *A note on the number of squares in a word*, Theoretical Computer Science, 380, 3 (2007), 373-376.

[6] B. Matschke, F. Santos, and C. Weibel, *The width of 5-prismatoids and smaller non-Hirsch polytopes* `http://www.cs.dartmouth.edu/~weibel/hirsch.php` (2011).

[7] F. Santos, *A counterexample to the Hirsch conjecture*, arXiv:1006.2814v1 (2010).