Discrete Applied Mathematics 163 (2014) 268–274

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

A d-step approach to the maximum number of distinct squares and runs in strings

Antoine Deza*, Frantisek Franek

Advanced Optimization Laboratory, Department of Computing and Software, McMaster University, Hamilton, Ontario, Canada

ARTICLE INFO

Article history:
Received 20 December 2011
Received in revised form 10 October 2013
Accepted 11 October 2013
Available online 30 October 2013

Keywords:
Strings
Runs
Distinct squares
(d, n)-parameterization
d-step approach

ABSTRACT

Fraenkel and Simpson conjectured in 1998 that the number of distinct squares in a string is at most its length. Kolpakov and Kucherov conjectured in 1999 that the number of runs in a string is also at most its length. Since then, both conjectures attracted the attention of many researchers and many results have been presented, including asymptotic lower bounds for both, asymptotic upper bounds for runs, and universal upper bounds for distinct squares in terms of the length. In this survey we point to the combined role played by the length and the number of distinct symbols of the string in both problems. Let us denote $\sigma_d(n)$, respectively $\rho_d(n)$, the maximum number of distinct primitively rooted squares, respectively runs, over all strings of length n containing exactly d distinct symbols. We study both functions $\sigma_d(n)$ and $\rho_d(n)$ and revisit earlier results and conjectures with the (d, n)-parameterized approach. The parameterized approach reveals regularities for both $\sigma_d(n)$ and $\rho_d(n)$ which have been computationally verified for all known values. In addition, the approach provides a computationally efficient framework. We were able to determine all previously known $\rho_2(n)$ values for $n \leq 60$ in a matter of hours, confirming the results reported by Kolpakov and Kucherov, and were able to extend the computations up to $n = 74$. Similarly, we were able to extend the computations up to $n = 70$ for $\sigma_2(n)$. We point out that $\sigma_2(33) < \sigma_3(33)$; that is, among all strings of length 33, no binary string achieves the maximum number of distinct primitively rooted squares, and that $\sigma_2(n) \leq 2n - 85$ for $n \geq 70$. The computations also reveal the existence of unexpected binary run-maximal string of length 66 containing a quadruple of identical symbols *aaaa*.

1. Introduction

A *square*, or a *tandem repeat* is a fundamental regularity in a string, and the simplest of *repetitions*. We denote this fact as u^2 indicating the concatenation of a string u with a copy of itself; u is referred to as the *generator* of the square and the length of u is referred to as the *period* of the square. A *primitively rooted square* is a square whose generator is *primitive*, i.e. not a repetition itself. Similarly, a *primitively rooted repetition* is a repetition whose generator is *primitive*. A *run*, a maximal primitively rooted repetition with a possibly fractional exponent, was conceptually introduced by Main in 1989 [20]. The term *run* was coined by Iliopoulos, Moore, and Smyth in 1997 [16]. A run in a string x encoded by a four-tuple (s, p, e, t) has a primitive *generator* $x[s \dots s + p]$ of length p repeating e times $(e \geq 2)$ followed by the prefix of the generator of length t $(0 \leq t < p)$. More precisely, $x[s + i] = x[s + i + rp]$ for $1 \leq i < s + p$ and $1 \leq r < e$, and $x[s + i] = x[s + i + rp]$ for $1 \leq i \leq t$ and $1 \leq r \leq e$. The maximality in this context means that the same is neither true for $s - 1$ nor for $s + 1$. Thus, the knowledge of all runs succinctly captures the knowledge of all occurrences of all repetitions. It is natural to ask about the maximum number of distinct squares or runs in a string and to expect both to depend primarily on the length of the string and, secondarily, on the number of distinct symbols.

* Corresponding author. Tel.: +1 905 525 9140.
E-mail addresses: deza@mcmaster.ca (A. Deza), franek@mcmaster.ca (F. Franek).

0166-218X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dam.2013.10.021

Table 1
$(d, n-d)$ table for $\sigma_d(n)$ with $2 \le d \le 15$ and $2 \le n-d \le 15$.

		$n-d$													
		2	3	4	5	6	7	8	9	10	11	12	13	14	15
d	2	**2**	2	3	3	4	5	6	7	7	8	9	10	11	12
	3	2	**3**	3	4	4	5	6	7	8	8	9	10	11	12
	4	2	3	**4**	4	5	5	6	7	8	9	9	10	11	12
	5	2	3	4	**5**	5	6	6	7	8	9	10	10	11	12
	6	2	3	4	5	**6**	6	7	7	8	9	10	11	11	12
	7	2	3	4	5	6	**7**	7	8	8	9	10	11	12	12
	8	2	3	4	5	6	7	**8**	8	9	9	10	11	12	13
	9	2	3	4	5	6	7	8	**9**	9	10	10	11	12	13
	10	2	3	4	5	6	7	8	9	**10**	10	11	11	12	13
	11	2	3	4	5	6	7	8	9	10	**11**	11	12	12	13
	12	2	3	4	5	6	7	8	9	10	11	**12**	12	13	13
	13	2	3	4	5	6	7	8	9	10	11	12	**13**	13	14
	14	2	3	4	5	6	7	8	9	10	11	12	13	**14**	14
	15	2	3	4	5	6	7	8	9	10	11	12	13	14	**15**

1.1. Distinct squares in strings

The problem of the number of distinct squares, when the types of the squares in a string are counted rather than their occurrences, was first introduced by Fraenkel and Simpson [12], showing that the number of distinct squares in a string of length n is at most $2n$. In particular, the number of primitively rooted distinct squares in strings of length n is bounded by $2n-8$ for $n \ge 5$, and, for binary strings of length n, by $2n-29$ for $n \ge 22$. Fraenkel and Simpson also gave an infinite sequence of strings of strictly increasing lengths with a number of primitively rooted distinct squares asymptotically approaching the strings length from below and conjectured that the number of distinct squares in a string is at most its length. Their work relied on an improved Lemma 10 of Crochemore and Rytter [7] stating that if u^2, v^2, w^2 are prefixes of a string x, w is primitive, and $|u| > |v| > |w|$, then $|u| \ge |v| + |w|$. Ilie [14] provided a simpler proof of the main lemma of [12] and presented an asymptotic upper bound of $2n - \Theta(\log n)$ in [15].

We focus on primitively rooted squares as opposed to all squares for the following reasons: conceptually it is closer to runs since they are primitively rooted too; Kubica et al. [19] showed that the number of non-primitively rooted distinct squares is bounded by $\lfloor \frac{n}{2} \rfloor - 1$, and computationally obtained values for both appear to be essentially the same. For the rest of the paper, the term *square* means a primitively rooted square unless explicitly stated otherwise.

1.2. Runs in strings

Though there may be as many as $O(n \log n)$ repetitions in a string of length n, see [5], it was hoped that the more succinct notation of runs would eliminate the need to list all the repetitions. Kolpakov and Kucherov [18] showed in 1999 that the number of runs in a string is $O(n)$ and conjectured that the maximum number of runs in a string is at most its length n. Several authors have presented asymptotic upper and lower bounds for the maximum number of runs over all strings of length n, see Crochemore and Ilie [6] for upper bounds, and Matsubara et al. [21] for lower bounds, and references in both.

2. Parameterized approach

In this survey we point out the importance of considering both the length n and the number d of distinct symbols in a string. We revisit earlier results and conjectures with this parameterized viewpoint. In particular, we hope to infer tighter upper bounds for the maximum number of distinct squares and runs in a string expressed in terms of d and n. A string x of length n with d distinct symbols is referred to as a (d, n)-string, $\boldsymbol{s}(x)$ denotes the number of distinct primitively rooted squares and $\boldsymbol{r}(x)$ denotes the number of runs in a string x. The symbol $\sigma_d(n)$ denotes the maximum number of distinct primitively rooted squares and $\rho_d(n)$ the maximum number of runs over all (d, n)-strings. A (d, n)-string x satisfying $\boldsymbol{s}(x) = \sigma_d(n)$ is referred to as a *square-maximal*, while a string x satisfying $\boldsymbol{r}(x) = \rho_d(n)$ is referred to as a *run-maximal* string.

We first discuss some elementary properties of the function $\sigma_d(n)$ whose values are presented in a so-called $(d, n-d)$ *table* where $\sigma_d(n)$ is the value on the d's row and the $(n-d)$'s column of the table, and point to ways of applying reductions to the problem of bounding the maximum number of distinct squares. The computed values with $2 \le d \le 15$ and $2 \le n-d \le 15$ of the $(d, n-d)$ table for $\sigma_d(n)$ are given in Table 1 with the main diagonal in bold. The up-to-date table of all computed values is available online at [10].

Then, we discuss some elementary properties of the function $\rho_d(n)$ which values are presented in a similar $(d, n-d)$ *table*, and similarly point to ways of applying reductions to the problem of bounding the maximum number of runs. The computed values with $2 \le d \le 15$ and $2 \le n-d \le 15$ of the $(d, n-d)$ table for $\rho_d(n)$ are given in Table 2 with the main diagonal in bold. The up-to-date table of all computed values is available online at [4].

While there are similarities, the investigation of distinct squares is different from the investigation of runs in many ways. For instance, while the concatenation of two strings may merge some runs from both strings, it would not merge distinct

Table 2
$(d, n-d)$ table for $\rho_d(n)$ with $2 \le d \le 15$ and $2 \le n-d \le 15$, and where $\rho_2(12) = 8$ and $\rho_3(15) = 10$ corresponding to the singularity (3,15) are in bold italic.

		$n-d$													
		2	3	4	5	6	7	8	9	10	11	12	13	14	15
d	2	**2**	2	3	4	5	5	6	7	8	***8***	10	10	11	12
	3	2	**3**	3	4	5	6	6	7	8	9	***10***	11	11	12
	4	2	3	**4**	4	5	6	7	7	8	9	10	11	12	12
	5	2	3	4	**5**	5	6	7	8	8	9	10	11	12	13
	6	2	3	4	5	**6**	6	7	8	9	9	10	11	12	13
	7	2	3	4	5	6	**7**	7	8	9	10	10	11	12	13
	8	2	3	4	5	6	7	**8**	8	9	10	11	11	12	13
	9	2	3	4	5	6	7	8	**9**	9	10	11	12	12	13
	10	2	3	4	5	6	7	8	9	**10**	10	11	12	13	13
	11	2	3	4	5	6	7	8	9	10	**11**	11	12	13	14
	12	2	3	4	5	6	7	8	9	10	11	**12**	12	13	14
	13	2	3	4	5	6	7	8	9	10	11	12	**13**	13	14
	14	2	3	4	5	6	7	8	9	10	11	12	13	**14**	14
	15	2	3	4	5	6	7	8	9	10	11	12	13	14	**15**

Table 3
$(d, n-2d)$ table for $\rho_d(n) - \sigma_d(n)$ with $2 \le d \le 7$ and $0 \le n-2d \le 17$, and where $\rho_2(13) - \sigma_2(13) = 0$ corresponding to the singularity (3, 15) is in bold italic.

		$n-2d$																	
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
d	2	0	0	0	1	1	0	0	0	1	***0***	1	0	0	0	1	1	2	1
	3	0	0	0	1	1	0	0	0	1	1	1	0	0	0	1	1	2	1
	4	0	0	0	1	1	0	0	0	1	1	1	0	0	0	1	1	2	1
	5	0	0	0	1	1	0	0	0	1	1	1	0	0	0	1	1	2	1
	6	0	0	0	1	1	0	0	0	1	1	1	0	0	0	1	1	2	1
	7	0	0	0	1	1	0	0	0	1	1	1	0	0	0	1	1	2	1

squares. On the other hand all the runs in both strings count, while the same is not true for distinct squares. The computed values of $\sigma_d(n)$ and $\rho_d(n)$ appear to be very close and hint at the simple relationship $\sigma_d(n) \le \rho_d(n)$ as illustrated in Table 3 where the entries are presented in a $(d, n-2d)$ table. The up-to-date table of all computed values $\rho_d(n) - \sigma_d(n)$ is available online at [10]. The computed values of $\sigma_d(n)$ and $\rho_d(n)$ lead to the hypothesized upper bounds:

$$\sigma_d(n) \le n - d - \lfloor \log_2 \lfloor (n+10-2d)/6 \rfloor \rfloor - \lceil \log_2 \lceil (n+3-2d)/5 \rceil \rceil \quad \text{for } n \ge 2d+2,$$
$$\rho_d(n) \le n - d - \lceil \log_2 \lceil (n+4-2d)/4 \rceil \rceil \quad \text{for } n \ge 2d.$$

Note that for $d = 2$ the upper bounds for $\sigma_2(n)$ and $\rho_2(n)$ are tight for, respectively, $6 \le n \le 19$ and $4 \le n \le 12$, and off by 1 for, respectively, $20 \le n \le 31$ and $13 \le n \le 37$. In addition, the upper bound for $\rho_3(n)$ is tight for $n = 6, \ldots, 16$. In terms of the maximum number $\sigma(n)$, respectively $\rho(n)$, of squares, respectively runs, over all strings of length n, the corresponding hypothesized upper bounds are:

$$\sigma(n) \le n - 2 - \lfloor \log_2 \lfloor (n+6)/6 \rfloor \rfloor - \lceil \log_2 \lceil (n-1)/5 \rceil \rceil \quad \text{for } n \ge 6,$$
$$\rho(n) \le n - 2 - \lceil \log_2 \lceil n/4 \rceil \rceil \quad \text{for } n \ge 4.$$

The values for $\sigma_d(n)$ and $\rho_d(n)$ computed to date indicate that $\sigma_d(n) - \sigma_{d-1}(n-2) = 1$ and $\rho_d(n) - \rho_{d-1}(n-2) = 1$ for $n \ge d \ge 3$ except for relatively rare pairs (n, d) satisfying either $\sigma_d(n) - \sigma_{d-1}(n-2) \ge 2$ or $\rho_d(n) - \rho_{d-1}(n-2) \ge 2$. For $\sigma_d(n)$, so far we have found two such pairs: (3, 35) as $\sigma_3(35) = 25$ and $\sigma_2(33) = 23$, and (3, 36) as $\sigma_3(36) = 26$ and $\sigma_2(34) = 24$, see [10]. For $\rho_d(n)$, so far we have found three such pairs: (3, 15) as $\rho_3(15) = 10$ and $\rho_2(13) = 8$, see Table 2 and the entries in bold italic, (3, 43) as $\rho_2(41) = 33$ and $\rho_3(43) = 35$, and (4, 44), as $\rho_3(42) = 33$ and $\rho_4(44) \ge \rho_3(43) = 35$, see [4]. We hypothesize that, though rare, there are infinitely many such pairs for both $\rho_d(n)$ and $\sigma_d(n)$. Our hypothesis implies that the values along a column of Table 3 are constant except for every such pair (d, n) and its corresponding entry in column $n - 2d$ and row $d - 1$ in the $(d, n-d)$ table. For illustration, for $\rho_d(n)$ and $(d, n) = (3, 15)$, the entry at column $n - d = 9$ and row $d = 2$ is depicted in bold italic in Table 3. This leads us to the following definition: we refer to a pair (d, n) such that either $\sigma_d(n) - \sigma_{d-1}(n-2) \ge 2$ or $\rho_d(n) - \rho_{d-1}(n-2) \ge 2$ as a *singularity*.

Remark 1. Though it is impossible to have 3 consecutive equal values in any row of the $(d, n-d)$ table for $\rho_d(n)$ as $\rho_d(n+2) > \rho_d(n)$, there is no such restriction for $\sigma_d(n)$. Such three times repeating values were found for binary strings of lengths 31, 32, and 33, however it is the only case known to us to date. Whenever $\sigma_d(n) = \sigma_d(n+1) = \sigma_d(n+2)$, it follows that $\sigma_{d+1}(n+2) > \sigma_d(n) = \sigma_d(n+2)$; that is, the maximum number of squares among all strings of length $n+2$ is not achieved by $(d, n+2)$-strings. In particular, since $\sigma_2(31) = \sigma_2(33) = \sigma_2(33) = 23$, any binary string of length 33 has at most 23 distinct squares while there is a ternary string with 24 distinct squares. For example, the following ternary string of length 33 has 24 distinct squares:*aababaababaabaababaabaababbabbacc*.

3. Basic properties of $\sigma_d(n)$ and $\rho_d(n)$

3.1. Basic properties of $\sigma_d(n)$

The following basic properties of $\sigma_d(n)$ were presented in [8,9] and are summarized in Proposition 2. The values of $\sigma_d(n)$ are increasing when moving right along a row of the $(d, n-d)$ table and the increase is of at most 2, the values are increasing when moving down along a column, the values are strictly increasing when moving along descending diagonals, the values under and on the main diagonal along a column are constant. In addition, the two values immediately above the main diagonal are equal and differ from the value on the main diagonal by at most 1 for $d \geq 4$. Note that the main diagonal of the $(d, n-d)$ table corresponds to the values of $\sigma_d(2d)$ for $d \geq 2$.

Proposition 2 (*[8,9]*).

(s_1) $0 \leq \sigma_d(n+1) - \sigma_d(n) \leq 2$ *for* $n \geq d \geq 2$,
(s_2) $\sigma_d(n) \leq \sigma_{d+1}(n+1)$ *for* $n \geq d \geq 2$,
(s_3) $\sigma_d(n) < \sigma_{d+1}(n+2)$ *for* $n \geq d \geq 2$,
(s_4) $\sigma_d(n) = \sigma_{d+1}(n+1)$ *for* $2d \geq n \geq d \geq 2$,
(s_5) $\sigma_d(n) \geq n - d$, $\sigma_d(2d+1) \geq d$ *and* $\sigma_d(2d+2) \geq d+1$ *for* $2d \geq n \geq d \geq 2$,
(s_6) $\sigma_{d-1}(2d-1) = \sigma_{d-2}(2d-2)$ *and* $0 \leq \sigma_d(2d) - \sigma_{d-1}(2d-1) \leq 1$ *for* $d \geq 4$,
(s_7) $1 \leq \sigma_{d+1}(2d+2) - \sigma_d(2d) \leq 2$ *for* $d \geq 2$.

Since $\sigma_2(70) = 55$ and $2 \geq \sigma_2(n+1) - \sigma_2(n) \geq 0$, we have the following slight improvement of the upper bound for $\sigma_2(n)$ in Corollary 3. In addition, $\sigma_{23}(46) = 23$ and $\sigma_{d+1}(2d+2) - \sigma_d(2d) \leq 2$ implies $\sigma_d(2d) \leq 2d - 23$ for $d \geq 23$. Thus, using $\sigma_d(n) \leq \sigma_{n-d}(2n-2d)$, the bound for the maximum number $\sigma(n)$ of distinct squares over all strings of length n can be slightly improved. We have (i) for $n - d \geq 23$: $\sigma_d(n) \leq \sigma_{n-d}(2n-2d) \leq 2n - 2d - 23$, and (ii) for $n - d \leq 23$: $\sigma_d(n) \leq \sigma_{n-d}(2n-2d) \leq n - d \leq 2n - d - 25$ for $n \geq 25$. We recall that the bounds given by Fraenkel and Simpson in [12] are $\sigma_2(n) \leq 2n - 29$ for $n \geq 22$, and $\sigma(n) \leq 2n - 8$ for $n \geq 5$.

Corollary 3. (c_1) $\sigma_2(n) \leq 2n - 85$ *for* $n \geq 70$,
(c_2) $\sigma(n) \leq 2n - 27$ *for* $n \geq 25$.

A *singleton* refers to a symbol in a string that occurs exactly once, while a *pair* refers to a symbol that occurs exactly twice. The following structural result for square-maximal strings on the main diagonal was noted in [8].

Proposition 4 (*[8]*). *Let d^* be the first integer, if it exists, such that $\sigma_{d^*}(2d^*) > d^*$, then any square-maximal $(d^*, 2d^*)$-string does not contain a pair and thus must contains at least $\lceil \frac{2d}{3} \rceil$ singletons.*

Propositions 2 and 4 yield Theorem 5 underlining the importance of the diagonals of the $(d, n-d)$ table with respect to the conjectured upper bound $n - d$ for $\sigma_d(n)$. In particular, Theorem 5 shows that in order to prove $\sigma_d(n) \leq n - d$ for all n and d it is enough to prove the bound for the special case $n = 2d$ for all d, i.e. for the main diagonal of the $(d, n-d)$ table. In other words, $\sigma_d(2d) \leq d$ for all d implies that the maximum number $\sigma(n)$ of runs over all strings of length n satisfies $\sigma(n) \leq n - 2$ for $n \geq 3$. This equivalence is further generalized to the special case $n = 4d$. In addition, the role played by $\sigma_d(2d)$ and $\sigma_d(2d+1)$ is underlined as well as the hypothesis that the square-maximal $(d, 2d)$-strings are, up to relabelling, unique.

Theorem 5 (*[8]*).

(e_1) $\{\sigma_d(n) \leq n - d$ *for* $n \geq d \geq 2\} \Longleftrightarrow \{\sigma_d(2d) \leq d$ *for* $d \geq 2\}$,
(e_2) $\{\sigma_d(n) \leq n - d$ *for* $n \geq d \geq 2\} \Longleftrightarrow \{\sigma_d(4d) \leq 3d$ *for* $d \geq 2\}$,
(e_3) $\{\sigma_d(n) \leq n - d$ *for* $n \geq d \geq 2\} \Longleftrightarrow \{\sigma_d(2d+1) - \sigma_d(2d) \leq 1$ *for* $d \geq 2\}$,
(e_4) $\{\sigma_d(2d+1) \leq d$ *for* $d \geq 2\} \Longrightarrow \{\sigma_d(n) < n - d$ *and* $\sigma_d(2d) = d$ *for* $n > 2d \geq 4\}$,
(e_5) $\{\sigma_d(2d) = \sigma_d(2d+1)$ *for* $d \geq 2\} \Longrightarrow \{\sigma_d(2d) = d$ *and* $\sigma_d(n) < n - d$ *for* $n > 2d \geq 4\}$,
(e_6) $\{\sigma_d(2d) = \sigma_d(2d+1)$ *for* $d \geq 2\} \Longrightarrow \{$ *square-maximal $(d, 2d)$-strings are, up to relabelling, unique and equal to* $a_1a_1a_2a_2 \ldots a_da_d\}$.

3.2. Basic properties of $\rho_d(n)$

The following basic properties of $\rho_d(n)$ are summarized in Proposition 6, see Section 4 for the proof. The values of $\rho_d(n)$ are increasing when moving right along a row of the $(d, n-d)$ table, the values are increasing when moving down along a column, the values are strictly increasing when moving along descending diagonals, the values under and on the main diagonal along a column are constant. In addition, the three values immediately above the main diagonal are equal and differ from the value on the main diagonal by at most 1 for $d \geq 5$. Note that the main diagonal of the $(d, n-d)$ table corresponds to the values of $\rho_d(2d)$ for $d \geq 2$.

Proposition 6. (r_1) *$\rho_d(n) \leq \rho_{d+1}(n+1)$ for $n \geq d \geq 2$,*
(r_2) *$\rho_d(n) \leq \rho_d(n+1)$ for $n \geq d \geq 2$,*
(r_3) *$\rho_d(n) < \rho_{d+1}(n+2)$ for $n \geq d \geq 2$,*
(r_4) *$\rho_d(n) = \rho_{d+1}(n+1)$ for $2d \geq n \geq d \geq 2$,*
(r_5) *$\rho_d(n) \geq n-d$, $\rho_d(2d+1) \geq d$ and $\rho_d(2d+2) \geq d+1$ for $2d \geq n \geq d \geq 2$,*
(r_6) *$\rho_{d-1}(2d-1) = \rho_{d-2}(2d-2) = \rho_{d-3}(2d-3)$ and $0 \leq \rho_d(2d) - \rho_{d-1}(2d-1) \leq 1$ for $d \geq 5$.*

Proposition 7 ([2]). *Let d^* be the first integer, if it exists, such that $\rho_{d^*}(2d^*) > d^*$, then any run-maximal $(d^*, 2d^*)$-string does not contain a symbol occurring exactly $2, 3, \ldots, 7$ or 8 times and thus must contain at least $\lceil \frac{7d}{8} \rceil$ singletons.*

Propositions 6 and 7 yield Theorem 8 underlining the importance of the diagonals of the $(d, n-d)$ table with respect to the conjectured upper bound $n-d$ for $\rho_d(n)$, see Section 4 for the proof. In particular, Theorem 8 shows that in order to prove $\rho_d(n) \leq n-d$ for all n and d it is enough to prove the bound for the special case $n = 2d$ for all d, i.e. for the main diagonal of the $(d, n-d)$ table. In other words, $\rho_d(2d) \leq d$ for all d implies that the maximum number $\rho(n)$ of runs over all strings of length n satisfies $\rho(n) \leq n-2$ for $n \geq 3$. This equivalence is further generalized to the special case $n = 9d$. In addition, the role played by $\rho_d(2d)$ and $\rho_d(2d+1)$ is underlined as well as the hypothesis that the run-maximal $(d, 2d)$-strings are, up to relabelling, unique.

Theorem 8. (e_1) *$\{\rho_d(n) \leq n-d$ for $n \geq d \geq 2\} \Longleftrightarrow \{\rho_d(2d) \leq d$ for $d \geq 2\}$,*
(e_2) *$\{\rho_d(n) \leq n-d$ for $n \geq d \geq 2\} \Longleftrightarrow \{\rho_d(9d) \leq 8d$ for $d \geq 2\}$,*
(e_3) *$\{\rho_d(n) \leq n-d$ for $n \geq d \geq 2\} \Longleftrightarrow \{\rho_d(2d+1) - \rho_d(2d) \leq 1$ for $d \geq 2\}$,*
(e_4) *$\{\rho_d(2d+1) \leq d$ for $d \geq 2\} \Longrightarrow \{\rho_d(2d) = d$ and $\rho_d(n) < n-d$ for $n > 2d \geq 4\}$,*
(e_5) *$\{\rho_d(2d) = \rho_d(2d+1)$ for $d \geq 2\} \Longrightarrow \{\rho_d(n) < n-d$ and $\rho_d(2d) = d$ for $n > 2d \geq 4\}$,*
(e_6) *$\{\rho_d(2d) = \rho_d(2d+1)$ for $d \geq 2\} \Longrightarrow \{$ square-maximal $(d, 2d)$-strings are, up to relabelling, unique and equal to $a_1a_1a_2a_2\ldots a_da_d\}$.*

Some hypothesized properties dealing with the maximal number of runs in a string can be restated in terms of the $(d, n-d)$ table. For example, the intuitive assumption that the number of runs cannot increase if the number of distinct symbol increases can be restated as: the values of the $(d, n-d)$ table cannot decrease along any counter-diagonal, that is $\rho_{d+1}(n) \leq \rho_d(n)$ for $n \geq d \geq 2$. In other words, the maximum along any counter-diagonal is achieved for $d = 2$, i.e. for binary strings.

Let us just remark that our approach is inspired by a similar $(d, n-d)$ table used for investigating the Hirsch bound for the diameter of polytopes. A *polyhedron* is an intersection of finitely many closed half-spaces, and a *polytope* is a bounded polyhedron. A (d, n)-polytope is a polytope of dimension d having n facets. The diameter $\boldsymbol{d}(P)$ of a polytope P is the smallest integer such any pair of vertices of P can be connected by an edge-path of length $\boldsymbol{d}(P)$ or less. Let $\Delta(d, n)$ denote the maximum possible diameter over all (d, n)-polytopes. The Hirsch conjecture, first posed in 1957 and stating that $\Delta(d, n) \leq n-d$, was disproved by Santos [22] in 2012 by exhibiting a violation on the main diagonal with $(d, n) = (43, 86)$. The associated Hirsch $(d, n-d)$ table exhibits similar regularities as the $(d, n-d)$ tables considered in this paper for $\sigma_d(n)$ and $\rho_d(n)$. Namely, it is known that $\Delta(d, n) \leq \Delta(d, n+1)$, $\Delta(d, n) < \Delta(d+1, n+2)$, and $\Delta(d, n) \leq \Delta(d+1, n+1)$ for $n \geq d \geq 2$; and that $\Delta(d, n) = \Delta(d+1, n+1)$ for $2d \geq n \geq d \geq 2$. In other words, if the values for $\Delta(d, n)$ are listed in a $(d, n-d)$ table where d is the index for the rows and $n-d$ the index for the columns, then the maximum of $\Delta(d, n)$ within a column is achieved on the main diagonal and all values below a value on the main diagonal are equal to that value. The role played by the main diagonal of the $(d, n-d)$ table was underlined in 1967 by Klee and Walkup [17] who showed the equivalency between the Hirsch conjecture and the d-step conjecture stating that $\Delta(d, 2d) \leq d$ for all $d \geq 2$. Note that the d-cube is a $(d, 2d)$-polytope having diameter d and therefore $\Delta(d, 2d) \geq d$ for any d. In other words, the string $a_1a_1a_2a_2\ldots a_da_d$ can be viewed as an analogue of the d-cube.

The value of $\Delta(d, n)$ provides a lower bound for the number of iterations of simplex methods for the worst case behavior. The simplex and central-path following primal–dual interior point methods are currently the most computationally successful algorithms for linear optimization. The curvature of a polytope, defined as the largest possible total curvature of the associated central path, can be regarded as the continuous analogue of its diameter. Considering the largest curvature $\Lambda(d, n)$, Deza et al. [11] proved the following continuous analogue of the equivalence between the Hirsch conjecture and the d-step conjecture: if $\Lambda(d, 2d) = \mathcal{O}(d)$ for all d, then $\Lambda(d, n) = \mathcal{O}(n)$.

3.3. Computational substantiation for tractable instances

The notion of an r-cover introduced in [2] was generalized in [3,9] to efficiently handle (n, d)-strings for the computation of both $\rho_d(n)$ and $\sigma_d(n)$. In the following definitions, a square is encoded as a pair (s, p) with s indicating the starting position and p indicating the period of the square. Note that the ending position of a square is $s+2p-1$ as we index strings starting with 0. Similarly, a run is encoded as a triple (s,e,p) where s is its starting position, e its ending position, and p its period.

Definition 9. An *r-cover* of a string x is a sequence of primitively rooted squares $\{S_i=(s_i,p_i) \mid 1\le i\le m\}$ such that

(1) none of the S_i's can be cyclically shifted to the left;
(2) $s_i<s_{i+1}\le s_i+2p_i<s_{i+1}+2p_{i+1}$ for $1\le i<m$, i.e. two consecutive squares are either adjacent or overlap without one containing the other;
(3) $\bigcup_{1\le i\le m} S_i=\bigcup_{1\le i\le m} x[s_i\ldots s_i+2p_i-1]=x$;
(4) for any run $R=(s,e,p)$ of x there is an $S_i=(s_i,p_i)$ containing the leading square (s,p) of R, i.e. $s_i\le s$ and $s+2p-1\le s_i+2p_i-1$.

Definition 10. An *s-cover* of a string x is a sequence of primitively rooted squares $\{S_i=(s_i,p_i) \mid 1\le i\le m\}$ such that

(1) $s_i<s_{i+1}\le s_i+2p_i$ and $s_i+2p_i-1<s_{i+1}+2p_{i+1}-1$ for $1\le i<m$, i.e. two consecutive squares are either adjacent or overlapping;
(2) $\bigcup_{1\le i\le m} S_i=\bigcup_{1\le i\le m} x[s_i\ldots s_i+2p_i-1]=x$;
(3) for any occurrence of a square $S=(s,p)$ in x, there is an S_i containing S, i.e. $s_i\le s$ and $s+2p-1\le s_i+2p_i-1$.

A straightforward heuristics to obtain an efficient lower bound $\sigma_2^-(n)$ for $\sigma_2(n)$ is proposed in [9]. Moreover, the value $\sigma_d^-(n)=\max\ \{\sigma_{d-1}(n-1),\ \sigma_{d-1}(n-2)+1,\ \sigma_d(n-1)\}$ is used as an efficient lower bound for $\sigma_d(n)$ for $d\ge 3$. In both cases, by *efficient* we mean the fact that $\sigma_d(n)-\sigma_d^-(n)\le 1$ for all pair (d,n) we have dealt with so far. Furthermore, as shown in [9], a square-maximal string with at least $\sigma_d^-(n)+1$ distinct squares must have a specific *s*-cover satisfying certain density conditions. Thus, a search for a string with at least $\sigma_d^-(n)+1$ distinct squares can be limited to such strings only, significantly reducing the search space and allowing the determination of $\sigma_d(n)$ for previously intractable values, see [10]. In a similar fashion, a straightforward heuristics to obtain an efficient lower bound $\rho_2^-(n)$ for $\rho_2(n)$ is given in [2] and the value $\rho_d^-(n)=\max\ \{\rho_{d-1}(n-1),\rho_{d-1}(n-2)+1,\rho_d(n-1)\}$ is used as an efficient lower bound for $\rho_d(n)$ for $d\ge 3$. Again, by *efficient* we mean the fact that $\rho_d(n)-\rho_d^-(n)\le 1$ for all pair (d,n) we have dealt with so far. Similarly, as shown in [2], a run-maximal string with at least $\rho_d^-(n)+1$ run must have a specific *r*-cover satisfying certain density conditions. Thus, as for run-maximality, a search for a string with at least $\rho_d^-(n)+1$ can be limited to such strings only, significantly reducing the search space and allowing the determination of $\rho_d(n)$ for previously intractable values, see [4]. Roughly measuring the efficiency of our approach and its implementation, it appears that we are able to handle strings of twice the length of strings tractable by the previous approaches. The properties presented in this paper were computationally checked for all currently known instances of n and d. The subroutine computing the number of distinct squares or runs in a string uses the C++ implementation of the algorithm introduced in [13].

4. Proofs of Proposition 6 and Theorem 8

4.1. Elementary lemma

We recall some basic observations. The simple proofs for the first 2 items are omitted while the third one is presented in [3].

Lemma 11. (l_1) *A symbol occurring exactly twice can occur in at most one run.*
(l_2) *If a run-maximal string consists of pairs, then each pair consists of* 2 *adjacent symbols.*
(l_3) *If a run-maximal* (d,n)*-string has a singleton, then* $\rho_d(n)=\rho_{d-1}(n-1)$ *or* $\rho_d(n)=\rho_d(n-1)$.

4.2. Proof of Proposition 6

Item (r_1). Consider a run-maximal (d,n)-string x. The string y obtained from x by appending a new symbol satisfies $\boldsymbol{r}(x)=\boldsymbol{r}(y)\le\rho_{d+1}(n+1)$, thus $\rho_d(n)\le\rho_{d+1}(n+1)$.
Item (r_2). Consider a run-maximal (d,n)-string x. The string y obtained from x by appending any of the d symbols satisfies $\boldsymbol{r}(x)\le\boldsymbol{r}(y)\le\rho_d(n+1)$, thus $\rho_d(n)\le\rho_d(n+1)$.
Item (r_3). Consider a run-maximal (d,n)-string x. The string y obtained from x by appending two copies of a new symbol satisfies $\boldsymbol{r}(x)+1=\boldsymbol{r}(y)\le\rho_{d+1}(n+2)$, thus $\rho_d(n)<\rho_{d+1}(n+2)$.
Items (r_4) and (r_6). See [1] for a proof.
Item (r_5). Since $\rho_2(4)=\rho_2(5)=2$ and $\rho_{d+1}(n+2)>\rho_d(n)$, we have $\rho_d(2d)\ge d$ and $\rho_d(2d+1)\ge d$. Since $\rho_d(n)=\rho_{n-d}(2n-2d)$ for $2d\ge n>d\ge 2$, we have $\rho_d(n)\ge n-d$ for $2d\ge n>d\ge 2$.

4.3. Proof of Theorem 8

Item (e_1). Since the left hand side trivially implies the right hand side, we need to prove the converse. By Proposition 6, the right hand side is equivalent to $\rho_d(2d) = d$ for $d \geq 2$ and, with $\rho_d(n) = \rho_{d+1}(n+1)$ for $n \leq 2d$ and $\rho_d(n) \leq \rho_{n-d}(2n-2d)$ for $n > 2d$, it gives $\rho_d(n) \leq n-d$ for $n \geq d \geq 2$.
Item (e_2). See [1] for a proof.
Item (e_3). Let d^* be the first integer, if it exists, such that $\rho_{d^*}(2d^*) > d^*$. If a run-maximal $(d^*, 2d^*)$-string has a singleton them $\rho_{d^*}(2d^*) = \rho_{d^*-1}(2d^*-1) \leq \rho_{d^*-1}(2d^*-2) = d^*-1$ or $\rho_{d^*}(2d^*) = \rho_{d^*}(2d^*-1) = \rho_{d^*-1}(2d^*-2) = d^*-1$ which contradicts $\rho_{d^*}(2d^*) > d^*$. Thus, by Lemma 11, any run-maximal $(d^*, 2d^*)$-string consists of pairs of adjacent symbols which contradicts $\rho_{d^*}(2d^*) > d^*$; that is, $\rho_d(2d)$ for $d \geq 2$ and thus $\rho_d(n) \leq n-d$ for $n \geq d \geq 2$.
Item (e_4). By Proposition 6 $\rho_d(2d+1) \leq d$ for $d \geq 2$ implies that $\rho_d(2d) = \rho_d(2d+1) = d$ for $d \geq 2$. In addition, $\rho_d(n) = \rho_{d+1}(n+1)$ for $n \leq 2d$ and $\rho_d(n) \leq \rho_{n-d}(2n-2d)$ for $n > 2d$ gives $\rho_d(n) < n-d$ for $n > 2d \geq 4$.
Item (e_5). Let d^* be the first integer, if it exists, such that $\rho_{d^*}(2d^*) > d^*$. If a run-maximal $(d^*, 2d^*)$-string has a singleton them $\rho_{d^*}(2d^*) = \rho_{d^*-1}(2d^*-1) = \rho_{d^*-1}(2d^*-2) = d^*-1$ or $\rho_{d^*}(2d^*) = \rho_{d^*}(2d^*-1) = \rho_{d^*-1}(2d^*-2) = d^*-1$ which contradicts $\rho_{d^*}(2d^*) > d^*$. Thus, by Lemma 11, any run-maximal $(d^*, 2d^*)$-string consists of pairs of adjacent symbols which contradicts $\rho_{d^*}(2d^*) > d^*$; that is, $\rho_d(2d) = d = \rho_d(2d+1)$ for $d \geq 2$ and $\rho_d(n) < n-d$ for $n > 2d \geq 4$.
Item (e_6). Assume that a run-maximal $(d, 2d)$-string has a singleton, then $\rho_d(2d) = \rho_{d-1}(2d-1) = \rho_{d-1}(2d-2) = d-1$ or $\rho_d(2d) = \rho_d(2d-1) = \rho_{d-1}(2d-2) = d-1$ by Lemma 11, which contradicts $\rho_d(2d) = d$. In addition, Lemma 11 implies that all pairs consist of adjacent symbols.

Acknowledgements

The authors would like to thank Maxime Crochemore, Jakub Radoszewski, and Jamie Simpson for helpful suggestions, and the referees for their comments that helped to improve the exposition. Thanks also to Andrew Baker, Mei Jiang, and Min Jing Liu for their help to computationally check some hypothesized properties for small instances. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada, MITACS, and by the Canada Research Chairs program, and made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET).

References

[1] A. Baker, A. Deza, F. Franek, A parameterized formulation for the maximum number of runs problem, in: J. Holub, B.W. Watson, J. Žd'árek (Eds.), Festschrift for Bořivoj Melichar, Czech Technical University, Prague, Czech Republic, 2012, pp. 102–117. Also available at: http://www.stringology.org/papers/Festschrift_BM70.pdf.

[2] A. Baker, A. Deza, F. Franek, On the structure of run-maximal strings, Journal of Discrete Algorithms 14 (2012) 10–14.

[3] A. Baker, A. Deza, F. Franek, A computational framework for determining run-maximal strings, Journal of Discrete Algorithms 20 (2013) 43–50.

[4] A. Baker, A. Deza, F. Franek, Run-maximal strings, http://optlab.mcmaster.ca/~bakerar2/research/runmax/index.html.

[5] M. Crochemore, An optimal algorithm for computing the repetitions in a word, Information Processing Letters 12 (1981) 297–315.

[6] M. Crochemore, L. Ilie, L. Tinta, The "runs" conjecture, http://www.csd.uwo.ca/faculty/ilie/runs.html.

[7] M. Crochemore, W. Rytter, Squares, cubes and time–space efficient strings searching, Algorithmica 13 (1995) 405–425.

[8] A. Deza, F. Franek, M. Jiang, A d-step approach for distinct squares in strings, in: R. Giancarlo, G. Manzini (Eds.), Proceedings of Combinatorial Pattern Matching—CPM 2011, in: Lecture Notes in Computer Science, vol. 6661, 2011, pp. 77–89.

[9] A. Deza, F. Franek, M. Jiang, A computational framework for determining square-maximal strings, in: J. Holub, J. Žd'árek (Eds.), Proceedings of Prague Stringology Conference 2012, Czech Technical University, Prague, Czech Republic, 2012, pp. 112–119.

[10] A. Deza, F. Franek, M. Jiang, Square-maximal strings, http://optlab.mcmaster.ca/~jiangm5/research/square.html.

[11] A. Deza, T. Terlaky, Y. Zinchenko, A continuous d-step conjecture for polytopes, Discrete and Computational Geometry 41 (2009) 318–327.

[12] A.S. Fraenkel, J. Simpson, How many squares can a string contain? Journal of Combinatorial Theory, Series A 82 (1998) 112–120.

[13] F. Franek, M. Jiang, C. Weng, An improved version of the runs algorithm based on Crochemore's partitioning algorithm, in: J. Holub, J. Žd'árek (Eds.), Proceedings of Prague Stringology Conference 2011, Czech Technical University, Prague, Czech Republic, 2011, pp. 98–105.

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

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

[16] C.S. Iliopoulos, D. Moore, W.F. Smyth, A characterization of the squares in a Fibonacci string, Theoretical Computer Science 172 (1997) 281–291.

[17] V. Klee, D.W. Walkup, The d-step conjecture for polyhedra of dimension $d < 6$, Acta Mathematica 117 (1967) 53–78.

[18] R. Kolpakov, G. Kucherov, Finding maximal repetitions in a word in linear time, in: Proceedings of the 1999 Symposium on Foundations of Computer Science, FOCS'99, New York, USA, 1999, pp. 596–604.

[19] M. Kubica, J. Radoszewski, W. Rytter, T. Waleń, On the maximum number of cubic subwords in a word, European Journal of Combinatorics 34 (2013) 27–37.

[20] M.G. Main, Detecting leftmost maximal periodicities, Discrete Applied Mathematics 25 (1989) 145–153.

[21] W. Matsubara, K. Kusano, A. Ishino, H. Bannai, A. Shinohara, Lower bounds for the maximum number of runs in a string, http://www.shino.ecei.tohoku.ac.jp/runs/.

[22] F. Santos, A counterexample to the Hirsch conjecture, Annals of Mathematics 176 (1) (2012) 383–412.