A d-step approach for distinct squares in strings

A. Deza, F. Franek, and M. Jiang

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

CPM 2011, Palermo, Italy 27-29 June 2011

Outline

1. Motivation and background
2. $(d, n-d)$ table
3. Basic properties of $(d, n-d)$ table
4. Main results
5. Structure of relatively short square-maximal strings
6. Conclusion and future research

Motivation and background

This is a problem of counting *distinct types* of squares in strings rather than their occurrences. Eg: `aabaab` contains 2 distinct squares: `aa` and `aabaab`. This is a part of general investigation of *periodicities* in strings.

Similar, but different from counting runs: the example above has 3 runs `aabaab` and 2 distinct squares `aa`, `aabaab`, while `ababa` has 1 run `ababa` and 2 distinct squares `ab`, `ba`.

The problem was introduced by Fraenkel and Simpson in 1998. They showed that:

- the number of distinct squares in a string of length n is bounded from above by $2n$
- there is a lower bound of $n-o(n)$ asymptomatically approaching n from below for primitively rooted squares (where the generator is *primitive*, i.e. not being a repetition)

The upper bound is based on an application of the three square lemma by Crochemore and Rytter (1995) to derive the fact that at most two rightmost occurrences of squares can start at the same point.

Ilie (2005) provided a simpler proof of the main lemma and slightly improved the upper bound to $2n - \Theta(\log n)$ in 2007.

It is believed, that the number of distinct squares is bounded by the length of the string.

There are no further results or improvements despite the fact that the gap between the hypothesized bound and the known bound is so large. The problem is considered hard as very little of the combinatorics of squares is known and even less of possible structure of square-maximal strings.

It is intuitively clear that the bigger the alphabet of a string, the smaller the number and types of periodicities. In this sense, the problem of periodicities is hardest for binary strings.

A trivial example: 2 distinct squares for a binary alphabet: `aabb`, while only 1 distinct square for a ternary alphabet: `aabc`.

However, this intuitive belief never formalized and/or proven.

Another motivation comes from investigation of maximum number of runs, in the form of a conjecture: for any n, there exists a binary string of length n that attains the maximum number of runs.

We embarked on a systematic investigation of the primitively rooted distinct squares problem in relation to the size of the alphabet, thus $\sigma_d(n) = \max \{ s(\boldsymbol{x}) : |\boldsymbol{x}| = n \,\&\, |\mathcal{A}(\boldsymbol{x})| = d \}$ *where* $s(\boldsymbol{x})$ *denotes the number of distinct squares in* $\boldsymbol{x}$ *and* $\mathcal{A}(\boldsymbol{x})$ *the alphabet of the string* $\boldsymbol{x}$*.*

These values are put in a table, but not in the usual manner of $\sigma_d(n)$ is the value in the d-th row and the n-th column.

Instead, we organize the table in a slightly different manner so that the value $\sigma_d(n)$ is placed in the d-th row and the $(n-d)$-th column.

$(d, n-d)$ table

		$n-d$										
		1	2	3	4	5	6	7	8	9	10	11
	1	1	1	1	1	1	1	1	1	1	1	$\sigma_1(12)$
	2	1	2	2	3	3	4	5	6	7	7	$\sigma_2(13)$
	3	1	2	3	3	4	4	5	6	7	8	$\sigma_3(14)$
	4	1	2	3	4	4	5	5	6	7	8	$\sigma_4(15)$
	5	1	2	3	4	5	5	6	6	7	8	$\sigma_5(16)$
d	6	1	2	3	4	5	6	6	7	7	8	$\sigma_6(17)$
	7	1	2	3	4	5	6	7	7	8	8	$\sigma_7(18)$
	8	1	2	3	4	5	6	7	8	8	9	$\sigma_8(19)$
	9	1	2	3	4	5	6	7	8	9	9	$\sigma_9(20)$
	10	1	2	3	4	5	6	7	8	9	10	$\sigma_{10}(21)$
	11	$\sigma_{11}(12)$	$\sigma_{11}(13)$	.	.	.	.	.	.	.	.	

The main diagonal, the second diagonal

There are several reasons for organizing the table in this unorthodox manner:

- The regularities the table exhibits point to several ways of possible induction
- Point to a proper upper bound: $\sigma_d(n) \leq n-d$
- Allow a reduction of the problem, as the behaviour of the items on the main diagonal determines the behaviour of all entries.

Basic properties of $(d, n-d)$ table

For any $2 \leq d \leq n$, it is proven that:

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

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

(c) $\sigma_d(n) < \sigma_{d+1}(n+2)$
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 $n \leq 2d$
the values under and on the main diagonal along a column are constant

(e) $\sigma_d(n) \geq n-d$ for $n \leq 2d$
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) $\sigma_d(2d) - \sigma_{d-1}(2d-1) \leq 1$
the difference between the value on the main diagonal and the value immediately above it is no more than 1

Main results

This sections contains several propositions that are equivalent with the conjectured upper bound for $\sigma_d(n) \leq n-d$.

Theorem (The main diagonal dominates)

$\sigma_d(n) \leq n-d$ holds true for all $2 \leq d \leq n$ iff $\sigma_d(2d) \leq d$ for every $d \geq 2$.

Theorem (If the main diagonal and the second one are "close")

$\sigma_d(n) \leq n-d$ holds true for all $2 \leq d \leq n$ iff $\sigma_d(2d+1)-\sigma_d(2d) \leq 1$ for every $d \geq 2$.

Theorem (If second diagonal bounded, a stronger upper bound)

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

Theorem (If the main diagonal and the second one are the same, a stronger upper bound)

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

- Thus, to prove the conjecture in general, it is sufficient to prove it for the main diagonal.
- The sets of square-maximal strings of length n and d distinct symbols are irregular and unstructured, while the square-maximal strings on the main diagonal are regular and structured -- see Mei's website `http://optlab.mcmaster.ca/~jiangm5/research/square.html`
- If the conjecture does not hold, a counterexample will more likely be found in the main diagonal
- In the next sections we will show a possible way to prove the conjecture for the main diagonal

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 (i.e. on the main diagonal or in its vicinity). The main result in this section requiring several lemmas can be summarized in the following lemma 8 and theorem 5.

Lemma (Lemma 8)

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

$S_d(n)$ *denotes the set of all strings of length n with d distinct symbols* ***singleton*** *is a symbol that occurs in a string just once.*

The theorem can be viewed as yet another reformulation of the conjecture:

Theorem (Theorem 5)

$\sigma_d(n) \leq n-d$ holds true for all $2 \leq d \leq n$ iff $\sigma_d(4d) \leq 3d$ for every $d \geq 2$.

Conclusion and future research

The lemmas and the theorem of the previous section all follow the same scenario: *we know the conjecture is satisfied up to column d and column d is investigated.*

Several lemmas are used to show that a square-maximal string $\boldsymbol{x} \in S_d(d)$ either obeys the conjecture, or cannot contain a *pair*. The hardest part of the proof of Lemma 8 is to show, that if it contains a *triple*, it is balanced by an existence of a unique 6 symbols, i.e. at least a *6-tuple* must exist.

It is quiet conceivable that the following can be proven: $\boldsymbol{x}$ either obeys the conjecture, or if it has an *k-tuple*, then it must contain an $(k+1)$-*tuple*. This would prove the non-existence of a counterexample to the conjecture on the main diagonal, and hence everywhere.

Let us remark that our approach was inspired by a similar $(d, n-d)$ table used for investigation of the Hirsch bound for the diameter of bounded polytopes. The table exhibits similar regularities as the $(d, n-d)$ table considered in this talk.

Hirsch conjecture – the edge-vertex graph of an n-facet polytope in d-dimensional Euclidean space has diameter no more than $n-d$ – was recently disproved by Santos by exhibiting a violation on the main diagonal with $d = 43$ which was further improved to $d = 20$ by Weibel.

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 recently by Baker, Deza, and Franek shows that the conjectured upper bound of $n-d$ for the number of runs is equivalent with the upper bound of $8d$ for strings in $S_d(9d)$ for every $d \geq 2$.

THANK YOU