

A d -step Approach for Distinct Squares in Strings

Mei Jiang

Joint work with Antoine Deza and Frantisek Franek

Department of Computing and Software
McMaster University

LSD & LAW, Feb 2011

Outline

- 1 Introduction
- 2 $(d, n - d)$ Table
- 3 Conjecture Reformulations
- 4 Relatively Short Square-Maximal Strings Structure
- 5 Conclusions

Outline

- 1 Introduction
- 2 $(d, n - d)$ Table
- 3 Conjecture Reformulations
- 4 Relatively Short Square-Maximal Strings Structure
- 5 Conclusions

Background

- In 1998 Fraenkel and Simpson showed the number of distinct squares in a string of length n is bounded from above by $2n$ and gave a lower bound asymptotically approaching n from below; for primitively rooted squares, they showed that an upper bound of $n - o(n)$ is valid for infinitely many values of n .
- In 2005 Ilie provided a simpler proof of Fraenkel and Simpson's 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.

d -step Approach

- We investigate the problem of distinct squares in relationship to the alphabet of the string.
- We construct a table whose rows are indexed by d and columns are indexed by $n - d$ with entries of $\sigma_d(n)$.
- We conjecture that the upper bound for the maximum number of primitively rooted distinct squares is $n - d$.
- d -step approach was inspired by the techniques used for investigating the Hirsch bound for the maximum possible diameter over all d -dimensional polytopes with n facets.

Basic Notation

- A **square** is a repetition with power of 2, **distinct squares** means only the types of the squares are counted, **primitively rooted distinct squares** means the generator itself is not a repetition.
- A **run**, a maximal fractional primitively rooted repetition, is formed by a maximal repetition followed by a tail.
- $s(\mathbf{x})$ denotes the number of primitively rooted distinct squares in a string x .
- $\sigma_d(\mathbf{n})$ denotes the maximum number of primitively rooted distinct squares over all strings of length n containing exactly d distinct symbols.
- A **singleton** refers to a symbol in a string that occurs exactly once, a **pair** occurs exactly twice, a **triple** occurs exactly three times, and in general an **k -tuple** (k times).

Outline

- 1 Introduction
- 2 (d, n - d) Table**
- 3 Conjecture Reformulations
- 4 Relatively Short Square-Maximal Strings Structure
- 5 Conclusions

$(d, n-d)$ Table Basic Properties

		$n-d$										
		1	2	3	4	5	6	7	8	9	10	...
d	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	...

$(d, n-d)$ Table: $\sigma_d(n)$ with $1 \leq d \leq 10$ and $1 \leq n-d \leq 10$

For all $n \geq d \geq 2$:

- 1 $\sigma_d(n) \leq \sigma_d(n+1)$
- 2 $\sigma_d(n) \leq \sigma_{d+1}(n+1)$
- 3 $\sigma_d(n) < \sigma_{d+1}(n+2)$
- 4 $\sigma_d(n) = \sigma_{d+1}(n+1)$ for $n \leq 2d$
- 5 $\sigma_d(n) \geq n-d$ for $n \leq 2d$
- 6 $\sigma_d(2d) - \sigma_{d-1}(2d-1) \leq 1$

Outline

- 1 Introduction
- 2 $(d, n - d)$ Table
- 3 Conjecture Reformulations**
- 4 Relatively Short Square-Maximal Strings Structure
- 5 Conclusions

Theorem 1

Theorem 1

For all $n \geq d \geq 2$, $\sigma_d(n) \leq n - d \iff \sigma_d(2d) = d$ for all $d \geq 2$

		$n - d$				
		d
d				
			
	d			d		
	...			d	...	
	...			d		...

Proof.

- $n < 2d$, constant under the diagonal.
- $n > 2d$, smaller or equal than the diagonal value.



Theorem 2

Theorem 2

For all $n \geq d \geq 2$, $\sigma_d(n) \leq n - d \iff \sigma_d(2d+1) - \sigma_d(2d) \leq 1$ for all $d \geq 2$

		$n - d$				
		...	$d-1$	d
d				
	$d-1$		$\sigma_{d-1}(2d-2)$	$\sigma_{d-1}(2d-1)$	}	=
	d		≤ 1	$\sigma_d(2d)$		
	

Proof.

d is the least s.t. $\sigma_d(2d) > d$.

Remove the singleton,

$\sigma_{d-1}(2d-1) = \sigma_d(2d)$.

$\sigma_d(2d) - \sigma_{d-1}(2d-2) \leq 1$, and

$\sigma_{d-1}(2d-2) = d-1$. Thus

$\sigma_d(2d) \leq d$. □

Theorem 3

Theorem 3

For all $d \geq 2$, if $\sigma_d(2d+1) \leq d$, then

- ① $\sigma_d(n) \leq n - d$ for all $n \geq d \geq 2$
- ② $\sigma_d(n) \leq n - d - 1$ for all $n > 2d \geq 4$

		$n - d$				
		...	d	$d+1$
d				
	d		d	d		
	$d+1$		d	$d+1$		
	...		d	$d+1$...	
	...		d	$d+1$...

Proof.

- $\sigma_d(2d) = \sigma_d(2d+1) = d$.
- $n > 2d$, smaller than the diagonal value.

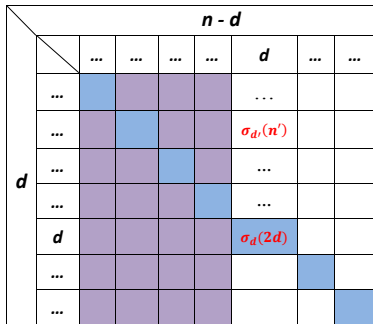


Outline

- 1 Introduction
- 2 $(d, n - d)$ Table
- 3 Conjecture Reformulations
- 4 Relatively Short Square-Maximal Strings Structure**
- 5 Conclusions

Relatively Short Square-Maximal Strings Structure

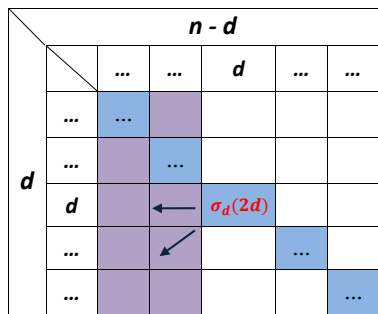
- We investigate the structure of square-maximal strings on the main diagonal.
 - If $\sigma_d(2d) = d$ then at least one of the square maximal string is in the form of $aabbccddeeff\dots$
 - If $\sigma_d(2d) > d$ then the square maximal string is a counterexample. We investigate its structure and draw conclusions for counterexamples with $n \leq 4d$.



Pairs

Lemma 1

Let d is the least s.t. for some x , $s(x) = \sigma_d(2d) > d$. Then x does not contain a pair.



Proof.

The pair: $x[i_0] = x[i_1] = C$.

- Occurs in only one square. Replace the first C with a new symbol \hat{C} .
 $d - 1 \geq \sigma_{d+1}(2d) \geq \sigma_d(2d) - 1$.
- Occurs in a non-trivial run $uvCwuvCwu$. Remove wuv between C 's.
 $d - k \geq \sigma_d(2d - k) \geq \sigma_d(2d) - k$,
 where $k = |w| + |u| + |v|$.



Triples

Lemma 2

Let d is the least s.t. for some x , $s(x) = \sigma_d(2d) > d$. Then x can only contain a triple $x[i_0] = x[i_1] = x[i_2] = C$ that satisfies:

- 1 $x[i_0]$ and $x[i_1]$ occur in a run $r_1 = u_1 v_1 C w_1 u_1 v_1 C w_1 u_1$, where $|u_1| \geq 1$,
- 2 $x[i_1]$ and $x[i_2]$ occur in a run $r_2 = u_2 v_2 C w_2 u_2 v_2 C w_2 u_2$, where $|u_2| \geq 1$, and where $i_1 - i_0 \neq i_2 - i_1$,
- 3 either $u_1 v_1$ is a proper suffix of $u_2 v_2$, or $w_2 u_2$ is a proper prefix of $w_1 u_1$.

Triples (cont.)

Proof.

$$r_1 : \quad u_1 v_1 C w_1 u_1 v_1 C w_1 u_1$$

$$r_2 : \quad u_2 v_2 C w_2 u_2 v_2 C w_2 u_2$$

- Show it is impossible to have only two symbols occur in a run.
- Show it is impossible to have three symbols occur in the same run.
- Show it is impossible to have both ends are “long”.



Singletons Estimation

Lemma 3

Let d is the least s.t. for some x , $s(x) = \sigma_d(2d) > d$. Then x has at least $\lceil \frac{2d}{3} \rceil$ singletons.

Proof.

- Let $u_1 v_1$ is a proper suffix of $u_2 v_2$, $a = u_1[0]$. a occurs at least 6 times in the r_1 and r_2 . We assign 5 a 's to the triple. It can be shown this assignment is mutually disjoint with others.

$$\begin{array}{l}
 r_1 : \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
 \quad \quad u_1 v_1 C w_1 u_1 v_1 C w_1 u_1 \\
 r_2 : \quad \quad \quad \quad u_2 v_2 C w_2 u_2 v_2 C w_2 u_2 \\
 \quad \quad \quad \quad \quad \quad \cdot \quad \cdot
 \end{array}$$

- m_0 : the number of triples, m_1 : the number of other multiply occurring symbols (at least 4 times), m_2 : the number of singletons.

$$2d \geq 8m_0 + 4m_1 + m_2 \tag{1}$$

$$d \leq 2m_0 + m_1 + m_2 \tag{2}$$

Thus, $m_2 \geq \lceil \frac{2d}{3} \rceil$

Theorem 5

Theorem 5

For all $n \geq d \geq 2$, $\sigma_d(n) \leq n - d \iff \sigma_d(4d) \leq 3d$ for all $d \geq 2$

		$n - d$						
		d
		
	...					$\sigma_{d'}(4d')$		
	$\lceil \frac{2d}{3} \rceil$	
		
d	d					$\sigma_d(2d)$		
	...							
	...							

Proof.

d is the least s.t. $\sigma_d(2d) > d$.

Remove $\lceil \frac{2d}{3} \rceil$ singletons.

$\sigma_{d'}(4d') \geq \sigma_d(2d) > d$ and

$3d' = d$. Thus, $\sigma_{d'}(4d') > 3d'$. \square

Theorem 5 (cont.)

- We construct $(d, n-3d)$ table, the conjectured upper bound is true if all the main diagonal entries satisfies $\sigma_d(4d) \leq 3d$.

		$n - 3d$						
		1	2	3	4	...	d	...
d	1	$\sigma_1(4)$						
	2		$\sigma_2(8)$					
	3			$\sigma_3(12)$				
	4				$\sigma_4(16)$			
		
	d						$\sigma_d(4d)$	

- In general term, $(d, n-kd)$ table may be constructed and the conjecture is equivalent with $\sigma_d((k + 1)d) \leq kd$ for all $d \geq 2$ on the main diagonal .






Outline

- 1 Introduction
- 2 $(d, n - d)$ Table
- 3 Conjecture Reformulations
- 4 Relatively Short Square-Maximal Strings Structure
- 5 Conclusions

Conclusions

- We exhibit the usefulness of investigating the main diagonal of $(d, n-d)$ table for tackling the conjectured upper bound.
 - To prove the conjecture by showing that the first counterexample has an impossible structure. i.e. it cannot contain an k -tuple, or if it contains an k -tuple, then it must contain another symbol with a frequency $> k$.
 - To disprove the conjecture by finding a counterexample on the diagonal.
- The Hirsch conjecture was recently disproved by Santos by exhibiting a violation on the diagonal with $d = 20$.
- Let's remark the techniques we used for "pushing up" the main diagonal can be applicable to the verification of the conjectured upper bound.

References

-  A. S. FRAENKEL and J. SIMPSON, *How Many Squares Can a String Contain?*, Journal of Combinatorial Theory Series A, 82, 1 (1998), 112-120.
-  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.
-  L. ILIE, *A note on the number of squares in a word*, Theoretical Computer Science, 380, 3 (2007), 373-376.
-  F. SANTOS, *A counterexample to the Hirsch conjecture*, arXiv:1006.2814v1 (2010).
-  A. DEZA, F. FRANEK, and M. JIANG, *A d -step approach for distinct squares in strings*, AdvOL Technical Report 2011/01, Dept. of Computing and Software, McMaster University, Canada.

THANK YOU!