Computational Substantiation of the *d*-step Conjecture for Distinct Squares Revisited

Frantisek Franek and Michael Liut

Department of Computing and Software
McMaster University, Hamilton, Ontario, Canada
&
Department of Mathematical and Computational Sciences
University of Toronto Mississauga, Mississauga, Ontario, Canada

PSC 2021, Czech Technical University, Prague
August 31, 2021

Outline

McMaster University
UNIVERSITY OF TORONTO MISSISSAUGA

Motivation and background

- In a pivotal paper in 1981, *Crochemore* showed that the maximum number of maximal repetitions in a string is $O(n \log(n))$, attained by Fibonacci strings.
- *Maximal repetitions*, a precursor to *runs*, may contain several squares bundled up, so bounding the maximum number of squares is a different problem.
- In 1998, in another pivotal paper, *Fraenkel* and *Simpson* showed that the number of occurrences of squares is bounded by $n \log_\Phi(n) \approx 1.441 n \log_2 2(n)$ (Φ denotes the golden ratio).
- Improved in 2020 by Bannai et. al to $n \log_2(n)$.

- The main result by *Fraenkel* and *Simpson* in their 1998 paper is that the maximum number of **distinct squares**, when types rather than occurrences are counted, is bounded by $2n$. They conjectured that the bound should be $\leq n$.
- The most significant aspect of their work was the fact that only 2 rightmost squares can start at the same position.
- The combinatorics analysis of so-called double squares was pioneered by *Lam*, and fully developed by *Deza*, *Franek*, and *Thierry* in 2015, giving an upper bound for MNDS[1] as $\frac{11}{6}n \approx 1.83n$.

McMaster University
UNIVERSITY OF TORONTO MISSISSAUGA

[1] MNDS=maximum number of distinct squares

- Since then, several partial results concerning the densities of distinct squares distribution had been published – *Blanchet-Sadri et. al* and *Manea et. al*.
- In 2011, *Deza*, *Franek*, and *Jiang* presented the d-step approach to the problem and conjectured the bound to be $\leq n-d$ where d is the number of distinct symbols in the string (d-step conjecture).
- In 2012 they introduced *a computational substantiation of the d-step approach to the number of distinct squares problem* that allowed to approximately double the length of the strings for computational verification of MNDS conjecture and d-step conjecture.

- Note that the problem has two versions – counting all distinct squares, or a simpler version of counting all distinct *primitively rooted* squares.
- Fraenkel+Simpson's and Deza+Franek+Thierry's results are for all distinct squares, while Deza+Franek+Jiang were formulated for all distinct primitively rooted squares.
- There does not seem to be any essential reason not to be able to reformulate Deza+Franek+Jiang's result for all distinct squares, however the posted results are for primitively rooted version as the software used counted only the primitively rooted squares.

$(d, n{-}d)$ table

A string $\boldsymbol{x}$ is a (d, n)-string if it has d distinct symbols and its length equals n

abba is a $(2, 4)$-string (the distinct symbols are *a* and *b*)

abcabdbabd is a $(4, 10)$-string (the distinct symbols are *a*, *b*, *c*, and *d*)

$s(\boldsymbol{x}) =$ *the number of distinct squares in string* $\boldsymbol{x}$

$$\sigma_d(n) = \max\{s(\boldsymbol{x}) \ : \ \boldsymbol{x} \text{ is a } (d, n)\text{-string}\}$$

The $(d, n-d)$ table is an infinite table that contains the values $\sigma_d(n)$ for all $n \geq 2$ and all $2 \leq d \leq n$.

Normally, it would be expected to be organized in rows indexed by d and columns indexed by n:

		n					
		2	3	4	....	k	....
d	2			$\sigma_2(4)$			
	3						
	4						
	$\vdots$						
	r					$\sigma_r(k)$	
	$\vdots$						

In the $(d, n-d)$ table, we organize the entries differently, the rows are again indexed by d, but the columns are indexed by $n-d$:

	n-d						
d		2	3	4	....	k	....
	2			$\sigma_2(6)$			
	3						
	4						
	⋮						
	r					$\sigma_r(k+r)$	
	⋮						

n - d

d	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
2	2	2	3	3	4	5	6	7	7	8	9	10	11	12	12	13	13	14	15	16	17	18	19	20	20	21	22	23	23	23	24	25	26	27
3	2	3	3	4	4	5	6	7	8	8	9	10	11	12	13	13	14	14	15	16	17	18	19	20	21	21	22	23	24	24	25	26	26	27
4	2	3	4	4	5	5	6	7	8	9	9	10	11	12	13	14	14	15	15	16	17	18	19	20	21	22	22	23	24	25	25	26	27	
5	2	3	4	5	5	6	6	7	8	9	10	10	11	12	13	14	15	15	16	16	17	18	19	20	21	22	23	23	24	25	26			
6	2	3	4	5	6	6	7	7	8	9	10	11	11	12	13	14	15	16	16	17	17	18	19	20	21	22	23	24						
7	2	3	4	5	6	7	7	8	8	9	10	11	12	12	13	14	15	16	17	17	18	18	19	20	21	22	23	24	25					
8	2	3	4	5	6	7	8	8	9	9	10	11	12	13	13	14	15	16	17	18														
9	2	3	4	5	6	7	8	9	9	10	10	11	12	13	14	14	15	16	17	18	19													
10	2	3	4	5	6	7	8	9	10	10	11	11	12	13	14	15	15	16	17	18	19	20												
11	2	3	4	5	6	7	8	9	10	11	11	12	12	13	14	15	16	16	17	18	19	20	21											
12	2	3	4	5	6	7	8	9	10	11	12	12	13	13	14	15	16	17																
13	2	3	4	5	6	7	8	9	10	11	12	13	13	14	14	15	16	17	18															
14	2	3	4	5	6	7	8	9	10	11	12	13	14	14	15	15	16	17	18	19														
15	2	3	4	5	6	7	8	9	10	11	12	13	14	15	15	16	16	17	18	19	20													
16	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	17	18	19														
17	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	17	18	18															
18	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	18	19															
19	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20														
20	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	20														
21	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21														
22	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22													
23	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23												
24	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24											

d	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68
2	28	29	30	31	32	33	33	34	35	36	36	37	37	38	39	40	41	42	43	44	44	45	46	47	48	48	49	50	51	52	53	54	55
3	28	29	30	31	32	33	34																										
4																																	
5																																	
6																																	
7																																	
8																																	
9																																	
10																																	
11																																	
12																																	
13																																	
14																																	
15																																	
16																																	
17																																	
18																																	
19																																	
20																																	
21																																	
22																																	
23																																	
24																																	

Deza+Franek+Jiang showed that there are a lot of relationships in the table such as:

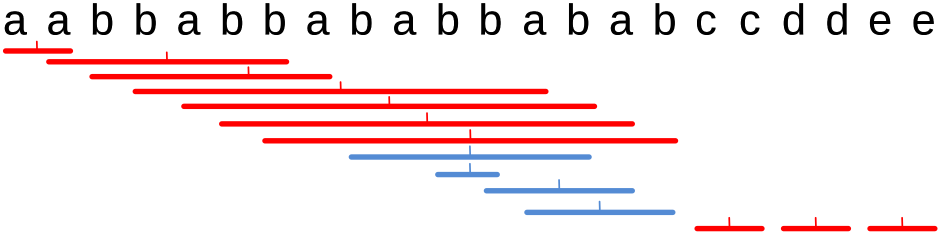
- Uniform below diagonal
  $(\forall\, 2 \le d)\big(\sigma_{d+k}(2d+k) = \sigma_d(2d)\big)$
- Diagonal rules $(\forall\, 2 \le d \le n)\big(\sigma_d(n) \le n-d\big) \Leftrightarrow (\forall\, 2 \le d)\big(\sigma_d(2d) = d\big)$
- Row increase
  $(\forall\, 2 \le d \le n)\big(\sigma_d(n) \le \sigma_d(n+1) \le \sigma_d(n)+2\big)$
- Column increase $(\forall\, 2 \le d \le n)\big(\sigma_{d+1}(n+1) \ge \sigma_d(n)\big)$
- Diagonal increase
  $(\forall\, 2 \le d)\big(\sigma_d(2d) \le \sigma_{d+1}(2d+2) \le \sigma_d(2d)+2\big)$

- The $(d, n-d)$ table entries can be filled in a fashion similar to dynamic programming, from left to right and top to bottom.
- This approach was utilized to compute the values.
- Knowing $\sigma_{d-1}(n-2)$, $\sigma_d(n-1)$, $\sigma_{d+1}(n)$, and $\sigma_{d-1}(n-1)$ strongly limits the pool of candidates for $\sigma_d(n)$.

	$\sigma_{d-1}(n-2)$	$\sigma_{d-1}(n-1)$
	$\sigma_d(n-1)$	$\sigma_d(n)$
	$\sigma_{d+1}(n)$	

S-cover

Some strings are “made of squares”:



aabbabbababbababccddee is one of the square-maximal (5,22)-strings; $\sigma_5(22) = 14$.

McMaster University

UNIVERSITY OF TORONTO MISSISSAUGA

Definition

A sequence $\{(a_i, b_i) : 1 \le i \le k\}$ is an **S-cover** of string $\boldsymbol{x} = \boldsymbol{x}[1..n]$, if

1. $(\forall i \in 1..k)$ $\boldsymbol{x}[a_i..b_i]$ is a rightmost square
2. $(\forall i \in 1..k-1)$ $a_i < a_{i+1}$ and $b_i < b_{i+1}$
3. $(\forall j \in 1..n)(\exists i \in 1..k)$ $a_i \le j \le b_i$
4. for every rightmost square (a, b) in $\boldsymbol{x}$, there is $i \in 1..k$ so that $a_i \le a < b \le b_i$

Observation

- *From condition 3, $a_1 = 1$ and $b_k = n$.*
- *If a string has an S-cover, then the S-cover is unique.*

Examples:

- the red squares form the S-cover of ***aabbababbababbababccddee***.
- the red squares + the first blue square is not an S-cover, the blue square violates the 2nd condition.
- ***abcabc*** has an S-cover, the S-cover consists of a single square *abcabc*.
- ***abbabbCabbabb*** does not have an S-cover, *C* is not in any rightmost square so any collection of rightmost squares cannot satisfy the 3rd condition.

- The $(d, n-d)$ table setup allows us to limit the search for square-maximal strings to strings with an S-cover.
- Note that strings with an S-cover are necessarily free of singletons (i.e., letters with a single occurrence).
- Since generating a square requires just to generate its root, it basically doubles the length of strings that can be processed.
- There are some additional constraints on the strings based on the properties of the $(d, n-d)$ table.

Generating counter-examples

- We changed the paradigm – instead of generating possible candidates for maximality, we try to generate counter-examples to the d-step conjecture, i.e. (d, n) strings with strictly more than $n-d$ rightmost squares.
- This change of paradigm significantly reduces the search space.
- The reduction of the search space is in the form of several conditions a possible counter-example must satisfy.
- It allowed us to extend the viable length of strings that can be processed, and hence extend the range of the validity of the d-step conjecture.

- Problem: since the result of computation for given d and n is an empty set, i.e. no counter-example had been generated, the result cannot be verified; the certificate is the code.
- But it is the same problem as with computing the square-maximal strings; the certificate is the code, as the maximality cannot be easily independently verified.
- As we keep discovering new necessary conditions that a counter-example must satisfy, the ultimate goal is to show analytically that counter-examples cannot exist.

The first lemma shows that induction on $n-d$ is well-founded and possible:

Lemma

Let $\boldsymbol{x}$ be a singleton-free (d, n)-string, $1 \le d < n$ and let d_1 be the number of distinct symbols in a non-empty proper prefix (resp. suffix) of $\boldsymbol{x}$ of length n_1. Then, $n_1 - d_1 < n - d$.

So, for all work we are assuming that the d-step conjecture holds for every $n_1 - d_1 < n - d$, and we are trying to generate counter-examples for d and n.

Necessary conditions for a (d, n)-string $\boldsymbol{x}$ to be a counter-example:

- It must have a special S-cover (and hence be singleton free).
- It must satisfy several density conditions.

The S-cover $\{(a_i, b_i) \ : \ 1 \le i \le k\}$ is **special** if:

- $k > 1$, so the S-cover has at least 2 squares.
- There is a square $(1, b)$ so that $b < b_1$, hence $(1, b_1)$ is an FS-double square.
  Note that $a_1 = 1$ *and that* $(1, b_1)$ *and* $(1, b)$ *are unique squares, i.e. both rightmost and leftmost occurrences.*
- The last square $(a_k, b_k) = (a_k, n)$ is a unique square.
- There is a unique square (a, n) so that $a_k < a$, hence $(1, a_k{-}n{+}1)$ is an FS-double square in $\boldsymbol{y}$, where $\boldsymbol{y} = \boldsymbol{x}[n]\boldsymbol{x}[n{-}1]...\boldsymbol{x}[2]\boldsymbol{x}[1]$, the string $\boldsymbol{x}$ in reverse.
- $\forall i \in 1..k{-}1$, $a_{i+1} < b_i$ and the intersection $\boldsymbol{x}[a_{i+1}..b_i]$ must contain all characters common to $\boldsymbol{x}[1..b_i]$ and $\boldsymbol{x}[a_{i+1}..n]$.

The density conditions are expressed using two arrays: $\boldsymbol{B}(\boldsymbol{i},\boldsymbol{j})$ is defined as the number of rightmost squares that start and $\boldsymbol{E}(\boldsymbol{i},\boldsymbol{j})$ that end in the interval $i..j$.

Let $1 \leq k < n$, let d_2 be the number of distinct symbols of $\boldsymbol{x}[k+1 \mathinner{..} n]$, and let e be the number of distinct symbols occurring in both $\boldsymbol{x}[1 \mathinner{..} k]$ and $\boldsymbol{x}[k+1 \mathinner{..} n]$.

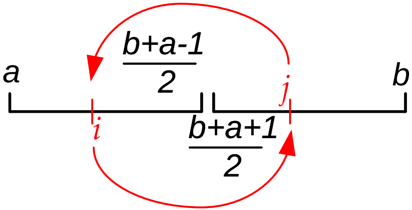
- $B(1,1) > 1$
- $B(1,2) > 2$
- $B(1,k) > k-d+d_2$
- $B(1,k)-E(1,k) > e$

In particular for binary strings:

- $B(1,k) > k$
- $B(1,k) - E(1,k) > 2$

An additional property that is not in the paper; let's define a symmetric and reflexive relation $\sim$ on $1..n$:

$i \sim j$ iff $i = j$ or there is a rightmost square (a, b) so that
$a \leq i \leq \frac{b+a-1}{2} < \frac{b+a+1}{2} \leq j \leq b$ and $i - a = j - \frac{b+a+1}{2}$ or
$a \leq j \leq \frac{b+a-1}{2} < \frac{b+a+1}{2} \leq i \leq b$ and $j - a = i - \frac{b+a+1}{2}$



Take the transitive closure of $\sim$. It is a relation of equivalence on $1..n$.

In simple terms, $i \sim j$ if $i = j$ or we can use successive rightmost squares to map i onto j.

Note that necessarily, if $i \sim j$, then $\boldsymbol{x}[i] = \boldsymbol{x}[j]$.

Necessary condition for the S-cover:

For any i in $1..a_{i+1}$ so that $\boldsymbol{x}[i]$ occurs in the intersection $\boldsymbol{x}[a_{i+1}..b_i]$, $i \sim j$ for some $j \in a_{i+1}..b_i$.

Conclusion

- We presented a set of necessary conditions for a counter-example for d-step conjecture for (d, n)-strings when the d-step conjecture is verified for all (d', n')-strings such that $n'-d' < n-d$.
- This allowed us to computationally verify the d-step conjecture for previously intractable lengths.
- The main goal is to discover more necessary conditions which would allow to prove the non-existence of counter-examples in an analytical way.

THANK YOU