On the Number of Distinct Squares

Frantisek (Franya) Franek

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

Invited talk - Prague Stringology Conference 2014

Outline

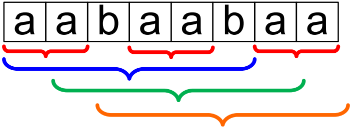
Introduction

objects of our research:

- finite strings over a finite alphabet $\mathcal{A}$
- required to have only $=$ and $\neq$ defined for elements of $\mathcal{A}$

what is the *maximum number of distinct squares* problem ?

counting types of squares rather than their occurrences:



6 occurrences of squares, but 4 distinct squares:
aa, *aabaab*, *abaaba*, and *baabaa*

research of periodicities is an active field

a deceptively similar problem of determining the maximum number of runs

occurrences of maximal (fractional) repetitions are counted

shown recently using the notion of Lyndon roots by Bannai et al. to be ***bounded by the length of the string***

Basic concepts

x is primitive $\iff x \neq y^p$ for any string y and any integer $p \geq 2$

Ex: *aab aab* is not primitive while *aabaaba* is

primitive root of x: the smallest y s.t. $x = y^p$ for some integer $p \geq 1$ *(is unique and primitive)*

u^2 is primitively rooted $\iff u$ is a primitive string

x and y are *conjugates* if $x = uv$ and $y = vu$ for some u, v

$x \lhd y \iff x$ is a proper prefix of y (i.e. $x \neq y$)

at most $O(n \log n)$ distinct squares

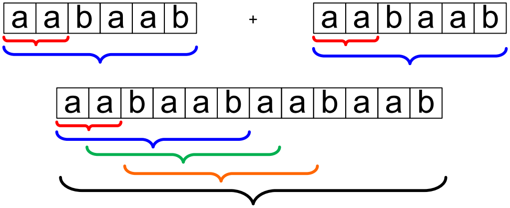
at most $O(\log n)$ squares can start at the same position

a	a	b	a	a	b	a	a	b	a	a	b

could it be $O(n)$? what would be the constant?

why this is not simple?

easy to compute for short strings, why not recursion?



concatenation

"destroys" multiply occurring existing types (aa, aabaab)

"creates" new types (abaaba, baabaa)

History

1994 *Fraenkel* and *Simpson*
How Many Squares Must a Binary Sequence Contain?
45 citations

what is the value of $g(k) =$ the longest binary word containing at most k distinct squares?

$g(0) = 3$, $g(1) = 7$, $g(2) = 18$ and $g(k) = \infty, k \geq 3$

motivated by the classic problem of combinatorics on words going all the way back to *Thue*: avoidance of patterns
an infinite ternary word avoiding squares

Fraenkel and Simpson introduced the term distinct squares for different types (or *shapes*) of squares

significant part of the paper – a construction of an infinite binary word containing only 3 distinct squares

focused on binary words as *Thue*'s result made the question irrelevant for larger alphabets

natural inversion of the question for all finite alphabets:

what is a number of distinct squares in a word ?

1998 *Fraenkel and Simpson* provided first non-trivial upper bound in How Many Squares Can a String Contain?
77 citations

Theorem

There are at most $2n$ distinct squares in a string of length n.

- *count only the rightmost occurrences*
- *show that if there are three rightmost squares $u^2 \lhd v^2 \lhd w^2$, then w^2 contains a farther occurrence of u^2*

based on Crochemore and Rytter 1995 Lemma: $|w| \geq |u| + |v|$

1998 *Fraenkel and Simpson* provided first non-trivial upper bound in How Many Squares Can a String Contain?
77 citations

Theorem

There are at most $2n$ distinct squares in a string of length n.

- *count only the rightmost occurrences*
- *show that if there are three rightmost squares $u^2 \lhd v^2 \lhd w^2$, then w^2 contains a farther occurrence of u^2*

based on Crochemore and Rytter 1995 Lemma: $|w| \geq |u| + |v|$

Crochemore and *Rytter*: *$u^2 \lhd v^2 \lhd w^2$ all primitively rooted, then $|u| + |v| \leq |w|$*

u^2 substring of the first $w \Rightarrow u^2$ substring of the second $w \Rightarrow$ u^2 cannot be rightmost

however u^2, v^2 and w^2 are rightmost and not primitively rooted

checking the details of the *Crochemore* and *Rytter*'s proof, *Fraenkel* and *Simpson* noted that only the primitiveness of u needed

most of their proof is thus devoted to the case when u^2 is not primitively rooted

Bai, Deza, F. (2014) generalization:

$u^2 \lhd v^2 \lhd w^2$, then either

(a) $|u| + |v| \leq |w|$

or *(inclusive or)*

(b) *u, v, and w have the same primitive root*

Fraenkel and *Simpson*'s result follows directly from it
(*u, v, and w have the same primitive root $\Rightarrow u^2$ not rightmost*)

2005 *Ilie* simpler proof not using *Crochemore* and *Rytter*'s lemma (*almost proved the generalized lemma*)

Fraenkel and Simpson further hypothesized that

$$\sigma(n) < n$$

$\sigma(n) = \max \{ \ s(x) \ : \ x \textit{ is a string of length } n \ \}$

$s(x)$ = *number of distinct squares in* x

and gave an infinite sequence of strings $\{x_n\}_{n=1}^{\infty}$ s.t.

$$|x_n| \nearrow \infty \quad \text{and} \quad \frac{s_p(x_n)}{|x_n|} \nearrow 1$$

$s_p(x)$ = *number of distinct primitively rooted squares in* x

2007 *Ilie* gives an asymptotic upper bound $2n - \theta(\log n)$

key idea – the last rightmost square of x must start way before the last position of x:

we saw this picture before: reversing it yields $\theta(\log\ n)$

$$\underbrace{\underbrace{\underbrace{\boxed{a}\boxed{a}}\boxed{b}\boxed{a}\boxed{a}\boxed{b}}\boxed{a}\boxed{a}\boxed{b}\boxed{a}\boxed{a}\boxed{b}}$$

2011 *Deza and F.* proposed a d-step approach and conjectured

$$\sigma_d(n) \leq n - d$$

$\sigma_d(n) = \max \{ s(x) \ : \ |x| = n \text{ with } d \text{ distinct symbols} \}$

- addresses dependence of the problem on the size of the alphabet
- is amenable to computational induction

up-to-date table of determined values:

`http://optlab.mcmaster.ca/~jiangm5/research/square.html`

d \ n - 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	36
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	28
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	28
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												

Lam (2013, preprint 2009)

claimed a universal upper bound $\sigma(n) \leq \frac{95}{48}n \approx 1.98n$

obtained by bounding # double squares and assuming at most single square everywhere else

double square = *pair of rightmost squares starting at the same position*

bounding of # double squares based on a complete taxonomy of mutual configurations of 2 or more double squares

is the taxonomy complete and sound ?

Deza, F., and Thierry

in How many double squares can a string contain?
(*to appear in Discrete Applied Mathematics*)

followed *Lam*'s approach

to further investigate the ***structural*** and ***combinatorial***

properties of double squares, resulting in

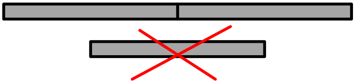
$$\sigma(n) \leq \frac{11}{6}n \approx 1.83n$$

presented today

Basic notions and tools

Lemma (*Synchronization Principle*)

Given a primitive string x, a proper suffix y of x, a proper prefix z of x, and $m \geq 0$, there are exactly m occurrences of x in yx^mz.



Lemma (*Common Factor Lemma*)

For any strings x and y, if a non-trivial power of x and a non-trivial power of y have a common factor of length $|x|+|y|$, then the primitive roots of x and y are conjugates.

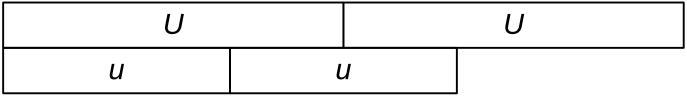
In particular, if x and y are primitive, then x and y are conjugates.

Note that both x and y must repeat at least twice

really a folklore, but proofs given in Two squares canonical factorization, PSC 2014, by *Bai, F., and Smyth*

Double squares

a configuration of two ***proportional*** squares u^2 and U^2



has been investigated in many different contexts:

- *Smyth et al.*: with intention to find a position for amortization argument for runs conjecture
- in computational framework by *Deza-F.-Jiang*: such configurations are used in *Liu*'s PhD thesis to speed up computation of $\sigma_d(n)$
- *Lam*: two rightmost squares form a particular structure

hence the following notation

- a *double square* (u, U) in a string x is a configuration of two squares u^2 and U^2 in x starting at the same position where $|u| < |U|$
- a double square (u, U) in x is *balanced* in x if u and U are *proportional*, i.e. $|U| < 2|u|$
- a balanced double square (u, U) in x is *factorizable* if either u is primitive, or U is primitive, or u^2 is rightmost in U^2
- a double square (u, U) in x is a *FS-double square* in x if both u^2 and U^2 are the rightmost occurrences in x

- (u, U) is a double square (balanced, factorizable) in x & x a factor in $y \Rightarrow$ double square (balanced, factorizable) in y

- (u, U) is a FS-double square in x & x a factor in $y \Rightarrow$ may not be FS-double square in y, ***unless*** x ***is a suffix of*** y

- (u, U) is a FS-double square in $x \Rightarrow$ balanced in x
  (if (u, U) not balanced, then u^2 is factor of U and hence a farther occurrence of u^2)

- (u, U) a FS-double square in $x \Rightarrow$ factorizable in x

Lemma

For a factorizable double square (u, U) there is a unique primitive string $\boldsymbol{u_1}$, a unique non-trivial proper prefix $\boldsymbol{u_2}$ of u_1, and unique integers $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$ satisfying $1 \leq e_2 \leq e_1$ such that $\boldsymbol{u = {u_1}^{e_1} u_2}$ and $\boldsymbol{U = {u_1}^{e_1} u_2 {u_1}^{e_2}}$.

a generalization by Bai, F., Smyth will be presented in a contributed talk

notation: $(u, U : u_1, u_2, e_1, e_2)$

$\overline{u}_2$ is defined as a suffix of u_1 so that $u_1 = u_2\overline{u}_2$

$\widetilde{u}_1 = \overline{u}_2 u_2$

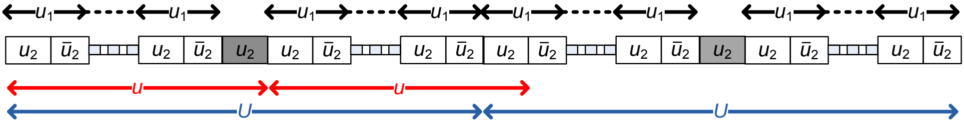
for a factorizable double square $\mathcal{U} = (u, U : u_1, u_2, e_1, e_2)$ simplified notation:

e_1 is denoted as $\mathcal{U}(1)$ end e_2 is denoted as $\mathcal{U}(2)$

Ex: for a factorizable double square $\mathcal{V}$, the ***shorter square*** is v^2, the ***longer square*** is V^2, and $(v, V : v_1, v_2, \mathcal{V}(1), \mathcal{V}(2))$, $v_1 = v_2\overline{v}_2$ and $\widetilde{v}_1 = \overline{v}_2 v_2$

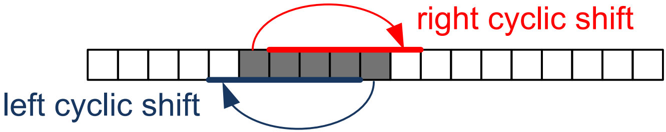
Structure of a factorizable double square

$$u_1{}^{\mathcal{U}(1)}\ u_2\ u_1{}^{\mathcal{U}(2)}\ u_1{}^{\mathcal{U}(1)}\ u_2\ u_1{}^{\mathcal{U}(2)}$$



only strings of length at least 10 may contain a factorizable double square:

$$|U^2| = 2((\mathcal{U}(1)+\mathcal{U}(2))|u_1|+|u_2|) \geq 2((1+1)2+1) = 10$$



right cyclic shift is determined by $lcp(u_1, \widetilde{u}_1)$

left cyclic shift is determined by $lcs(u_1, \widetilde{u}_1)$

lcp = largest common prefix
lcs = largest common suffix

$$u_1 = aaabaa,\ u_2 = aaab,\ \overline{u}_2 = aa,\ \mathcal{U}(1) = \mathcal{U}(2) = 2$$

```
aaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaa

[                           ] [                    ) (    ]                                            )
aaabaaaaabaaaaabaaabaaaaaabaaaaabaaaaabaaaaabaaabaaaaabaa...
  [                           ] [                    ) (    ]                                            )
.aabaaaaabaaaabaaabaaaaabaaaaabaaaabaaabaaaaabaaa..
    [                           ] [                    ) (    ]                                            )
..abaaaaabaaaabaaabaaaaaabaaaaabaaaabaaabaaaaabaaaa.
      [                           ] [                    ) (    ]                                            )
...baaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaabaaabaaaaabaaaaa
```

$u_1 = aaabaa$, $u_2 = aaab$, $\overline{u}_2 = aa$, $\mathcal{U}(1) = 2$, and $\mathcal{U}(2) = 1$

aaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaabaaaaa

[ ][ )( ])

aaabaaaaabaaaaabaaabaaaaaabaaaaabaaaabaaabaa...

[ ][ )( ])

.**aabaaaaabaaaaabaaabaaaaaabaaaaabaaaabaaabaa**..

[ ][ )( ])

..**abaaaaabaaaaabaaabaaaaaabaaaaabaaaabaaabaa**.

[ ][ )( ])

...**baaaaabaaaaabaaabaaaaaabaaaaabaaaabaaabaa**

Inversion factors

Definition

For a double square $\mathcal{U}$, $\bar{v}vv\bar{v}$ where $|\bar{v}| = |\bar{u}_2|$ and $|v| = |u_2|$ is an *inversion factor*

$$\mathcal{U} = {u_1}^{\mathcal{U}(1)} u_2 {u_1}^{\mathcal{U}(2)+\mathcal{U}(1)} u_2 {u_1}^{\mathcal{U}(2)} =$$

$${u_1}^{(\mathcal{U}(1)-1)} u_2 \bar{u}_2 u_2 u_2 \bar{u}_2 {u_1}^{\mathcal{U}(2)+\mathcal{U}(1)-2} u_2 \bar{u}_2 u_2 u_2 \bar{u}_2 {u_1}^{(\mathcal{U}(2)-1)}$$

N_1 N_2

natural inversion factors

a cyclic shift of an inversion factor is an inversion factor

determined by $lcp(u_1, \widetilde{u}_1)$ and $lcs(u_1, \widetilde{u}_1)$

```
                         L1                                          L2
[                          ][             )(                  ]                          )
aaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaabaaaabaa
aaabaaaaabaaaaabaaaaabaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaabaaaabaa
aaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaabaa
aaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaabaa
                            R1                                           R2
```

all inversion factors are cyclic shifts of the natural ones:

Lemma (*Inversion factor lemma*)

Given a factorizable double square $\mathcal{U}$, there is an inversion factor of $\mathcal{U}$ within the string U^2 starting at position $i \iff i \in [L_1, R_1] \cup [L_2, R_2]$.

FS-double squares

Theorem (*Fraenkel and Simpson, Ilie*)

At most 2 rightmost squares can start at the same position.

assume 3 rightmost squares $u^2 \lhd U^2 \lhd v^2$
(u, U) *is a factorizable double square so* $(u, U : u_1, u_2, e_1, e_2)$
first v *contains an inversion factor, so second* v *must also contain an inversion factor*
if it were from $[L_2, R_2]$*, then* $|v| = |U|$*, a contradiction*
so $u_1{}^{\mathcal{U}(1)} u_2 u_1{}^{\mathcal{U}(1)+\mathcal{U}(2)-1} u_2$ *must be a prefix of* v
v^2 *contains another occurrence of* $u_1{}^{\mathcal{U}(1)} u_2 u_1{}^{\mathcal{U}(1)} u_2 = u^2$,
a contradiction

FS-double squares

Theorem (*Fraenkel and Simpson, Ilie*)

At most 2 rightmost squares can start at the same position.

assume 3 rightmost squares $u^2 \lhd U^2 \lhd v^2$
(u, U) *is a factorizable double square so* $(u, U : u_1, u_2, e_1, e_2)$
first v *contains an inversion factor, so second* v *must also contain an inversion factor*
if it were from $[L_2, R_2]$*, then* $|v| = |U|$*, a contradiction*
so $u_1{}^{\mathcal{U}(1)} u_2 u_1{}^{\mathcal{U}(1)+\mathcal{U}(2)-1} u_2$ *must be a prefix of* v
v^2 *contains another occurrence of* $u_1{}^{\mathcal{U}(1)} u_2 u_1{}^{\mathcal{U}(1)} u_2 = u^2$*,*
a contradiction

key observation:

Lemma

Let x be a string starting with a FS-double square $\mathcal{U}$. Let $\mathcal{V}$ be a FS-double square with $\mathbb{s}(\mathcal{U}) < \mathbb{s}(\mathcal{V})$, then either

(a) $\mathbb{s}(\mathcal{V}) < R_1(\mathcal{U})$, *in which case either*

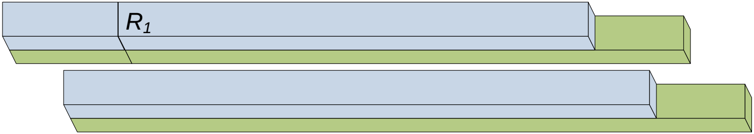
- (a_1) $\mathcal{V}$ *is an* α*-mate of* $\mathcal{U}$ *(cyclic shift), or*
- (a_2) $\mathcal{V}$ *is a* β*-mate of* $\mathcal{U}$ *(cyclic shift of* U *to* V*), or*
- (a_3) $\mathcal{V}$ *is a* γ*-mate of* $\mathcal{U}$ *(cyclic shift of* U *to* v*), or*
- (a_4) $\mathcal{V}$ *is a* δ*-mate of* $\mathcal{U}$ *(big tail),*

or

(b) $R_1(\mathcal{U}) \leq \mathbb{s}(\mathcal{V})$, *then*

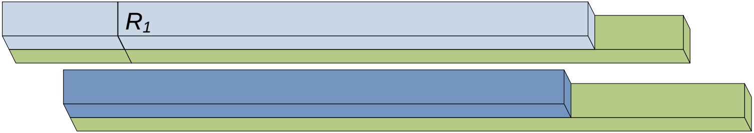
- (b_1) $\mathcal{V}$ *is a* ε*-mate of* $\mathcal{U}$ *(big gap).*

α-mate (cyclic shift):




```
[                    ][     )(           ]                )
aaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaabaa...
 [                    ][     )(           ]                )
.aabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaabaa..
```

β-mate (cyclic shift of U to V)



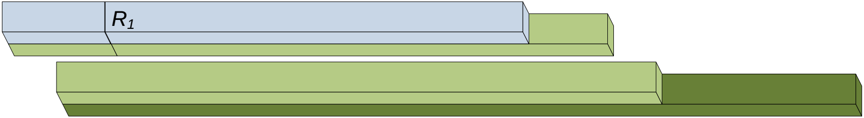
[ ] [ ) ( ])

baaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaa....

[ ] [ ) ( ])

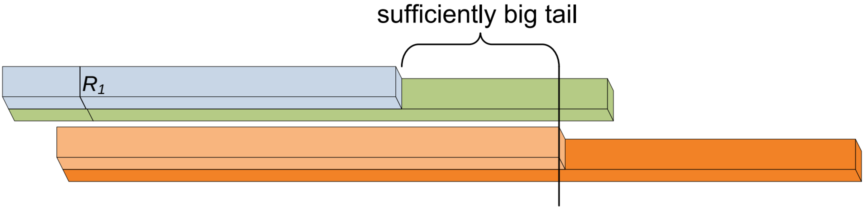
...aaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaaabaa.

γ-mate (cyclic shift of U to v)



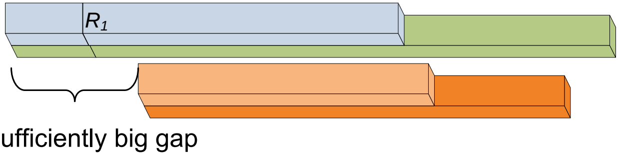

```
[              ][ )(           ]     )
aabaabaabaabaabaaabaabaabaabaabaabaaab
      [                ][       )(      ]                  )
      aabaabaabaaabaabaabaabaabaabaaabaabaaabaabaabaabaabaaba
```

δ-mate (big tail)




```
[              ][ )(            ]     )
aabaabaabaabaabaaabaabaabaabaabaabaaab
      [                  ][                  )(         ]                                    )
      aabaabaabaaabaabaabaabaabaabaaaabaabaabaaabaabaabaabaabaabaaabaabaabaabaabaabaaaabaabaabaaabaabaab
```

ε-mate (big gap)




```
[                  ][  )(              ]       )
aabaabaabaabaabaaabaabaabaabaabaabaaab
                    [                    ][     )(               ]              )
                    aabaabaabaabaabaaabaabaabaabaabaabaabaabaaabaab
```

FS-double squares: upper bound

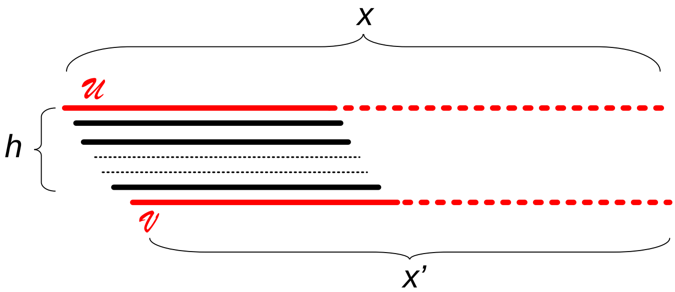
we show by induction that # FS-double squares $\delta(x)$ satisfies

$$\delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u|$$

where u^2 is the shorter square of the leftmost FS-double square of x

the fundamental observation lemma basically states that either

- δ-mate and gap G is “big", or
- ε-mate and tail T is “big", or
- α-mate or β-mate or γ-mate



Lemma (Gap-Tail lemma)

$\delta(x') \leq \frac{5}{6}|x'| - \frac{1}{3}|v|$ *implies*

$\delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u| + h - \frac{1}{2}|G(\mathcal{U}, \mathcal{V})| - \frac{1}{3}|T(\mathcal{U}, \mathcal{V})|$

we deal with α-mates, β-mates, and γ-mates in a special way

which is possible as they form families

- α-family, or
- $\alpha+\beta$-family, or
- $\alpha+\beta+\gamma$-family

$\mathcal{U}$-family consists only of α-mates

illustration of an α-family with $\mathcal{U}(1) = \mathcal{U}(2)$

```
aaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaa

[              ][          )(  ]                       )
aaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaa...
 [              ][          )(  ]                       )
.aabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaa..
  [              ][          )(  ]                       )
..abaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaa.
   [              ][          )(  ]                       )
...baaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaa
```

illustration of an α-family with $\mathcal{U}(1) > \mathcal{U}(2)$

aaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaabaaaaa

[ ][ )( ])
aaabaaaaabaaaaabaaabaaaaaabaaaaabaaaaabaaabaa . . .
[ ][ )( ])
. aabaaaaabaaaaabaaabaaaaaabaaaaabaaaaabaaabaaa . .
[ ][ )( ])
. . abaaaaabaaaaabaaabaaaaaabaaaaabaaaaabaaabaaaa .
[ ][ )( ])
. . . baaaaabaaaaabaaabaaaaaabaaaaabaaaaabaaabaaaaa

easy to bound the size of α-family, as it is determined by $lcp(u_1, \widetilde{u}_1)$: $|\alpha\text{-family}| \leq |u_1|$

- either there are no other FS-double squares, and then it can be shown directly that the bound holds, or

- there is a $\mathcal{V}$ underneath:
  $\mathcal{V}$ must be either γ-mate, or δ-mate, or ε-mate, and the Gap-Tail lemma can be applied to propagate the bound

$\mathcal{U}$-family consists of α-mates and β-mates

illustration of an $\alpha{+}\beta$-family

```
aaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaabaaaaa

[                                ][    )(                          ]           )
aaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaa..............α-segment starts
 [                                ][    )(                          ]           )
.aabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaa.............
  [                                ][    )(                          ]           )
..abaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaa............
   [                                ][    )(                          ]           )
...baaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaa...........
      [                          ][          )(              ]                       )
......aaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaa.........β-segment starts
       [                          ][          )(              ]                       )
.......aabaaaaabaaaaabaaaaabaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaa........
        [                          ][          )(              ]                       )
........abaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaa.......
         [                          ][          )(              ]                       )
.........baaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaa......
            [                    ][                )(   ]                                  )
...........aaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaabaa...β-segment starts
             [                    ][                )(   ]                                  )
............aabaaaaabaaaaabaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaabaaa..
              [                    ][                )(   ]                                  )
.............abaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaabaaaa.
               [                    ][                )(   ]                                  )
..............baaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaabaaaaa
```

more complicated to bound the size of $\alpha{+}\beta$-family:

$$|\alpha{+}\beta\text{-family}| \leq \begin{cases} \lceil \frac{\mathcal{U}(1)-\mathcal{U}(2)}{2} \rceil |u_1| & \text{if } \mathcal{U}(2) = 1 \\ \frac{\mathcal{U}(1)-\mathcal{U}(2)}{2} |u_1| & \text{if } \mathcal{U}(2) > 1 \end{cases}$$

- either there are no other FS-double squares, and then it can be shown directly that the bound holds, or
- there is a $\mathcal{V}$ underneath:
  $\mathcal{V}$ must be either δ-mate, or ε-mate, and the Gap-Tail lemma can be applied to propagate the bound *Special care needed for ε-mate case and super-ε-mate must be put in play !*

$\mathcal{U}$-family consists of α-mates, β-mates, and γ-mates

illustration of an $\alpha+\beta+\gamma$-family

```
              R1
[             ][ )(          ]     )     type
aabaabaabaabaabaaabaabaabaabaabaabaaab    5  1     <--- start of α-segment
 [             ][ )(          ]     )
 abaabaabaabaabaaabaabaabaabaabaabaaaba    5  1     <--- end of α-segment
   [          ][   )(    ]           )
   aabaabaabaabaaabaabaabaabaabaabaaabaab    4  2       <--- start of β-segment
    [          ][   )(    ]           )
    abaabaabaabaaabaabaabaabaabaabaaabaaba    4  2       <--- end of β-segment
      [                ][        )(      ]                   )
      aabaabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaaba   3  3    <--- start of γ-segment
       [                ][        )(      ]                   )
       abaabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaa   3  3
        [                ][        )(      ]                   )
        baabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaaa   3  3
         [                ][        )(      ]                   )
         aabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaaab   2  4 not a double square
          [                ][        )(      ]                   )
          abaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaaaba   2  4 not a double square
           [                ][        )(      ]                   )
           baabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaaabaa   2  4  not a double square
            [                ][        )(      ]                   )
            aabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaaabaab   2  4  not a double square
             [                ][        )(      ]                   )
             abaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaaabaaba   1  5  not a double square
              [                ][        )(      ]                   )
              baaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaaabaabaa   1  5  not a double square
              R1
```

it is quite complex to bound the size of $\alpha{+}\beta{+}\gamma$-family:

$$|\alpha{+}\beta{+}\gamma\text{-family}| \leq \tfrac{2}{3}(\mathcal{U}(1) + 1)|u_1|$$

- either there are no other FS-double squares, and then it can be shown directly that the bound holds, or
- there is a $\mathcal{V}$:
  $\mathcal{V}$ must be either δ-mate, or ε-mate, and the Gap-Tail lemma can be applied to propagate the bound

Main results

Theorem

There are at most $\lfloor 5n/6 \rfloor$ FS-double squares in a string of length n.

Corollary

There are at most $\lfloor 11n/6 \rfloor$ distinct squares in a string of length n.

Conclusion

- We presented a universal upper bound of $\frac{11n}{6}$ for the maximum number of distinct squares in a string of length n
- A universal upper bound of $\frac{5n}{6}$ for the maximum number of FS-double squares in a string of length n
- It improves the universal bound of $2n$ by *Fraenkel* and *Simpson*
- It improves the asymptotic bound of $2n - \Theta(\log n)$ by Ilie
- The combinatorics of double squares is interesting on its own and may be applicable to other problems

THANK YOU

A. Deza and F. Franek
A d-step approach to the maximum number of distinct squares and runs in strings
Discrete Applied Mathematics, 2014

A. Deza, F. Franek, and A. Thierry
How many double squares can a string contain?
to appear in Discrete Applied Mathematics

A. Deza, F. Franek, and M. Jiang
A computational framework for determining square-maximal strings
Proceedings of the Prague Stringology Conference, 2012

A.S. Fraenkel and J. Simpson
How Many Squares Must a Binary Sequence Contain?
The Electronic Journal of Combinatorics, 1995

A.S. Fraenkel and J. Simpson
How many squares can a string contain?
Journal of Combinatorial Theory, Series A, 1998

F. Franek, R.C.G. Fuller, J. Simpson, and W.F. Smyth
More results on overlapping squares
Journal of Discrete Algorithms, 2012

L. Ilie
A simple proof that a word of length n has at most $2n$ distinct squares
Journal of Combinatorial Theory, Series A, 2005

L. Ilie
A note on the number of squares in a word
Theoretical Computer Science, 2007

- E. Kopylova and W.F. Smyth
  The three squares lemma revisited
  Journal of Discrete Algorithms, 2012
- N. H. Lam
  On the number of squares in a string
  AdvOL-Report 2013/2, McMaster University, 2013
- M. J. Liu
  Combinatorial optimization approaches to discrete problems
  PhD thesis, Department of Computing and Software, McMaster University, 2013