Twins in words

M. Axenovich[1] Y. Person[2] S. Puzynina[3]

[1]Iowa State University, U.S.A. and Karlsruher Institut für Technologie, Germany

[2]Freie Universität Berlin, Institut für Mathematik, Germany

[3]University of Turku, Finland, and Sobolev Institute of Mathematics, Novosibirsk, Russia

Twins

- $S = s_1 \dots s_n$ a word of length n
- A (scattered) *subword* of S is a word $S' = s_{i_1} s_{i_2} \dots s_{i_l}$, where $i_1 < i_2 < \cdots < i_l$.
- *disjoint subwords* $s_{i_1} s_{i_2} \dots s_{i_l}$ and $s_{j_1} s_{j_2} \dots s_{j_t}$: $\{i_1, \dots, i_l\} \cap \{j_1, \dots, j_t\} = \emptyset$.

Definition 1

Twins: two disjoint identical subwords of S.

- $S = s_1 \dots s_n$ a word of length n
- A (scattered) *subword* of S is a word $S' = s_{i_1} s_{i_2} \dots s_{i_l}$, where $i_1 < i_2 < \dots < i_l$.
- *disjoint subwords* $s_{i_1} s_{i_2} \dots s_{i_l}$ and $s_{j_1} s_{j_2} \dots s_{j_t}$: $\{i_1, \dots, i_l\} \cap \{j_1, \dots, j_t\} = \emptyset$.

Definition 1

Twins: two disjoint identical subwords of S.

Example 2

$S = s_1 s_2 s_3 s_4 s_5 s_6 s_7 = 0001010$

Twins

- $S = s_1 \dots s_n$ a word of length n
- A (scattered) *subword* of S is a word $S' = s_{i_1} s_{i_2} \dots s_{i_l}$, where $i_1 < i_2 < \dots < i_l$.
- *disjoint subwords* $s_{i_1} s_{i_2} \dots s_{i_l}$ and $s_{j_1} s_{j_2} \dots s_{j_t}$: $\{i_1, \dots, i_l\} \cap \{j_1, \dots, j_t\} = \emptyset$.

Definition 1

Twins: two disjoint identical subwords of S.

Example 2

$S = s_1 s_2 s_3 s_4 s_5 s_6 s_7 = 0001010$

$S_1 = s_1 s_2 s_4 = 001$ and $S_2 = s_3 s_5 s_6 = 001$ are twins.

Twins

- $S = s_1 \dots s_n$ a word of length n
- A (scattered) *subword* of S is a word $S' = s_{i_1} s_{i_2} \dots s_{i_l}$, where $i_1 < i_2 < \dots < i_l$.
- *disjoint subwords* $s_{i_1} s_{i_2} \dots s_{i_l}$ and $s_{j_1} s_{j_2} \dots s_{j_t}$: $\{i_1, \dots, i_l\} \cap \{j_1, \dots, j_t\} = \emptyset$.

Definition 1

Twins: two disjoint identical subwords of S.

Example 2

$S = s_1 s_2 s_3 s_4 s_5 s_6 s_7 = 0001010$
$S_1 = s_1 s_2 s_4 = 001$ and $S_2 = s_3 s_5 s_6 = 001$ are twins.
$S'_1 = s_1 s_4 s_5 = 010$ and $S'_2 = s_2 s_6 s_7 = 010$ are also twins.

Twins

Problem

How large could the twins be in any word over a given alphabet?

Twins

Problem

How large could the twins be in any word over a given alphabet?

More precisely:

$f(S)$: the largest integer m such that there are twins of length m

In the example $S = 0001010$ one has $f(S) = 3$.

Twins

Problem

How large could the twins be in any word over a given alphabet?

More precisely:

$f(S)$: the largest integer m such that there are twins of length m

In the example $S = 0001010$ one has $f(S) = 3$.

We are interested in

$$f(n, \Sigma) = \min\{f(S) : S \in \Sigma^n\}$$

Twins

Problem

How large could the twins be in any word over a given alphabet?

More precisely:

$f(S)$: the largest integer m such that there are twins of length m

In the example $S = 0001010$ one has $f(S) = 3$.

We are interested in

$$f(n, \Sigma) = \min\{f(S) : S \in \Sigma^n\}$$

Trivial lower bound

$$f(n, \{0, 1\}) \geq \lfloor (1/3)n \rfloor$$

Twins

Problem

How large could the twins be in any word over a given alphabet?

More precisely:

$f(S)$: the largest integer m such that there are twins of length m

In the example $S = 0001010$ one has $f(S) = 3$.

We are interested in

$$f(n, \Sigma) = \min\{f(S) : S \in \Sigma^n\}$$

Trivial lower bound

$$f(n, \{0,1\}) \geq \lfloor (1/3)n \rfloor$$

$S = 001\ 101\ 111\ 010$

twins equal to 0110: $S = 001\ 101\ 111\ 010$

Our main result is

Theorem 3

There exists an absolute constant C such that

$$\left(1 - C\left(\frac{\log n}{\log\log n}\right)^{-1/4}\right) n \leq 2f(n, \{0,1\}) \leq n - \log n.$$

Twins

Our main result is

Theorem 3

There exists an absolute constant C such that

$$\left(1 - C\left(\frac{\log n}{\log\log n}\right)^{-1/4}\right) n \leq 2f(n, \{0,1\}) \leq n - \log n.$$

i.e., a binary word of length n has twins of length $n/2 - o(n)$

k-twins

Definition 4

k-twins in $S \in \Sigma^*$: k disjoint identical subwords of S

$f(S,k)$: the largest m so that S contains k-twins of length m each

$$f(n,k,\Sigma) = \min\{f(S,k) : \; S \in \Sigma^n\}$$

Theorem 5

For any integer $k \geq 2$, and alphabet Σ, $|\Sigma| \leq k$,

$$\left(1 - C|\Sigma|\left(\frac{\log n}{\log\log n}\right)^{-1/4}\right) n \leq kf(n,k,\Sigma) \leq n - \log n.$$

ε-regular words

The *density* of the letter q in S: $d_q(S) = |S|_q/|S|$.
$S[i, i+m] = s_i s_{i+1} \cdots s_{i+m}$

Definition 6 (ε-regular word)

For a positive ε, $\varepsilon < 1/3$, call a word S of length n over an alphabet Σ ε-regular if for every i, $\varepsilon n + 1 \leq i \leq n - 2\varepsilon n + 1$ and every $q \in \Sigma$ it holds that

$$|d_q(S) - d_q(S[i, i + \varepsilon n - 1])| < \varepsilon.$$

Example 7

Word S of length $n = 60$, density $1/2$:

011101010101000101001100110101010100110101010101111000010100

is ε-regular for $\varepsilon = 1/5$

Verification by definition:

- consider factors of length $\varepsilon n = (1/5) \cdot 60 = 12$ starting at positions $13, 14 \ldots, 37$
- compare their densities with the density $1/2$ of S
  - $S' = S[13, 24] = 000101001100$, $d(S') = 8/12$, $|8/12 - 1/2| < \epsilon = 1/5$
  - $S'' = S[14, 25] = 001010011001$, $d(S'') = 7/12$, $|5/12 - 1/2| < 1/5$
  - etc.

ε-regular partition

$\mathcal{S} := (S_1, \ldots, S_t)$: a *partition* of S if $S = S_1 S_2 \ldots S_t$

Definition 8 (ε-regular partition)

A partition $\mathcal{S}$ is an ε-regular partition of a word $S \in \Sigma^n$ if

$$\sum_{\substack{i \in [t] \\ S_i \text{ is not } \varepsilon-\text{regular}}} |S_i| \leq \varepsilon n,$$

i.e., the total length of ε-irregular factors is at most εn.

Regularity lemma

Key lemma:

Lemma 9 (Regularity Lemma for Words)

For every ε, t_0 *and* n *such that* $0 < \varepsilon < 1/3$, $t_0 > 0$ *and* $n > n_0 = t_0 \varepsilon^{-\varepsilon^{-4}}$, *any word* $S \in \Sigma^n$ *admits an* ε*-regular partition into* t *parts with* $t_0 \leq t \leq T_0 = t_0 3^{1/\varepsilon^4}$.

twins in ε-regular word

$S =$	01110	11001	10110	11010	11100	10110	10101
$S' =$	0 0	11 1	0 0	11 1	00	11 1	
$S'' =$		00	1 11	0 0	111	0 0	1 1 1

Large alphabets and small k-twins

Theorem 10

For any integer $k \geq 2$, and alphabet Σ of size l, $|\Sigma| > k$,

$$\left(\frac{k}{|\Sigma|} - C|\Sigma| \left(\frac{\log n}{\log \log n} \right)^{-\frac{1}{4}} \right) n \leq kf(n, k, \Sigma) \leq n - \max\{\alpha n, \log n\},$$

where $\alpha \in [0, 1/k]$ is the solution of the equation $l^{-(k-1)\alpha} \alpha^{-k\alpha} (1 - k\alpha)^{k\alpha - 1} = 1$, whenever such solution exists and 0 otherwise.

Large alphabets and small k-twins

Theorem 10

For any integer $k \geq 2$, and alphabet Σ of size l, $|\Sigma| > k$,

$$\left(\frac{k}{|\Sigma|} - C|\Sigma|\left(\frac{\log n}{\log\log n}\right)^{-\frac{1}{4}}\right) n \leq kf(n,k,\Sigma) \leq n - \max\{\alpha n, \log n\},$$

where $\alpha \in [0, 1/k]$ is the solution of the equation $l^{-(k-1)\alpha}\alpha^{-k\alpha}(1-k\alpha)^{k\alpha-1} = 1$, whenever such solution exists and 0 otherwise.

$k = 2$, $l = 5$: $\alpha < 0.49 \Rightarrow$ no twins of length $n/2 - o(n)$

Summary

We studied the following

Question:

is it true that any given word of length n over alphabet Σ has k-twins of size $n(1 - o(1))/k$ each?

Informally: the k-twins cover almost all letters of the word

Summary

We studied the following

Question:

is it true that any given word of length n over alphabet Σ has k-twins of size $n(1-o(1))/k$ each?

Informally: the k-twins cover almost all letters of the word

We have shown that the answer is:

- YES for $k \geq |\Sigma|$
- NO for some pairs $(k, |\Sigma|)$ with $k < |\Sigma|$, the smallest such pair we know is $(k, |\Sigma|) = (2, 5)$

Summary

We studied the following

Question:

is it true that any given word of length n over alphabet Σ has k-twins of size $n(1 - o(1))/k$ each?

Informally: the k-twins cover almost all letters of the word

We have shown that the answer is:

- YES for $k \geq |\Sigma|$
- NO for some pairs $(k, |\Sigma|)$ with $k < |\Sigma|$, the smallest such pair we know is $(k, |\Sigma|) = (2, 5)$

Open question

Is it true for $(k, |\Sigma|) = (k, k + 1)$?
We do not know even for $(2, 3)$.

Reference

Maria Axenovich, Yury Person, Svetlana Puzynina: A regularity lemma and twins in words. J. Comb. Theory, Ser. A 120(4): 733-743 (2013)