A REGULARITY LEMMA AND TWINS IN WORDS

MARIA AXENOVICH, YURY PERSON, AND SVETLANA PUZYNINA

Abstract. For a word S, let $f(S)$ be the largest integer m such that there are two disjoint identical (scattered) subwords of length m. Let $f(n,\Sigma) = \min\{f(S) : S \text{ is of length } n, \text{ over alphabet } \Sigma\}$. Here, it is shown that

$$2f(n,\{0,1\}) = n - o(n)$$

using the regularity lemma for words. In other words, any binary word of length n can be split into two identical subwords (referred to as twins) and, perhaps, a remaining subword of length $o(n)$. A similar result is proven for k identical subwords of a word over an alphabet with at most k letters.

Keywords: sequence, subword, identical subwords, twins in sequences.

1. Introduction

An *alphabet* Σ is a finite non-empty set of symbols called *letters*. Here, an alphabet is typically $\Sigma = \{0, 1, \dots, |\Sigma| - 1\}$. A *word* of length n, $S = s_1 \dots s_n$ is a sequence of elements from the alphabet, i.e., $s_i \in \Sigma$, $i = 1 \dots, n$. We denote the length of the word S by $|S|$. A (scattered) *subword* of S is a word $S' = s_{i_1} s_{i_2} \dots s_{i_l}$, where $i_1 < i_2 < \dots < i_l$. This notion was largely investigated in combinatorics on words and formal languages theory with special attention given to counting subword occurrences, different complexity questions, the problem of reconstructing a word from its subwords (see, e.g., [5, 10, 11]). Two subwords $s_{i_1} s_{i_2} \dots s_{i_l}$ and $s_{j_1} s_{j_2} \dots s_{j_t}$ of a word $s_1 s_2 \dots s_n$ are *disjoint* if the index sets are disjoint, i.e., $\{i_1, \dots, i_l\} \cap \{j_1, \dots, j_t\} = \emptyset$. For a word S, let $f(S)$ be the largest integer m such that there are two disjoint identical subwords of S, each of length m. We call such subwords *twins*. For example, if $S = s_1 s_2 s_3 s_4 s_5 s_6 s_7 = 0001010$, then $S_1 = s_1 s_2 s_4$ and $S_2 = s_3 s_5 s_6$ are both equal to 001. On the other hand $S_1' = s_1 s_4 s_5$ and $S_2' = s_2 s_6 s_7$ are twins of S, both equal to 010. In this example $f(S) = 3$.

The question we are concerned with is "How large could the twins be in any word over a given alphabet?" One of the classical problems related to this question is the problem of finding longest subsequence common to two given sequences, see for example [4, 7, 13]. Indeed, if we consider two disjoint subwords S_1 and S_2 of a given word S and find a subword S_1' of S_1 and a subword S_2' of S_2 such that $S_1' = S_2'$, then these subwords correspond to twins in S. Optimizing over all such S_1 and S_2 gives the largest twins.

Denoting Σ^n the set of words of length n over the alphabet Σ, let

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

Date: October 11, 2012.

The research of the first author is supported in part by NSF grant DMS-0901008.

The research of the third author is supported in part by grant 251371 of the Academy of Finland.

Observe first, that $f(n,\{0,1\}) \geq \lfloor (1/3)n \rfloor$. Indeed, consider any $S \in \Sigma^n$ and split it into consecutive triples. For each triple we add one of the letters that are repeated to S_1 and the other one to S_2. For example, if $S = 001\ 101\ 111\ 010$ then there are twins S_1, S_2, each equal to 0 1 1 0: $S = \mathbf{0}\underline{0}1\ \mathbf{1}0\underline{1}\ \mathbf{1}\underline{1}1\ \mathbf{0}1\underline{0}$, here one word is marked bold, and the other is underlined. In fact, we can find much larger identical subwords in any binary word.

Our main result is

Theorem 1. *There exists an absolute constant C such that*

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

In the proof we shall employ a classical density increment argument successfully applied in combinatorics and number theory, see e.g. the survey of Komlós and Simonovits [8] and some important applications [6] and [12]. We first show that we can partition any word S into consecutive factors that look as if they were random in a certain weak sense (we call them ε-regular). These ε-regular words can be partitioned (with the exception of ε proportion of letters) into two identical subwords. By appending these together for every ε-regular word, we eventually obtain identical subwords of S whose lengths satisfies the claimed inequality.

We generalize the notion of two identical subwords in words to a notion of k identical subwords. For a given word S, let $f(S,k)$ be the largest m so that S contains k pairwise disjoint identical subwords of length m each. Finally, let

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

We show the following bounds.

Theorem 2. *For any integer $k \geq 2$, and alphabet Σ, $|\Sigma| \leq k$, there exists a constant C such that*

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

In case when k is smaller than the size of the alphabet, we have the following bounds.

Theorem 3. *For any integer $k \geq 2$, and alphabet Σ, $|\Sigma| > k$, there exists a constant C such that*

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

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

We shall sometimes refer to two disjoint identical subwords as *twins*, three disjoint identical subwords as 3-*twins*, k disjoint identical subwords as k-*twins*. We shall prove the regularity lemma for words (Lemma 6) in Section 2 and will prove Theorem 1 in Section 3. We shall prove Theorems 2, 3 in Section 4. We shall ignore any divisibility issues as these will not affect our arguments.

2. Definitions and Regularity Lemma for Words

For a (scattered) *subword* $S' = s_{i_1}s_{i_2}\dots s_{i_l}$, of a word $S = s_1s_2\dots s_n$, we call the set $\{i_1, i_2, \dots, i_l\}$ a *support* of S' in S, and write $\mathrm{supp}(S')$, so the length of S', $|S'| = |\mathrm{supp}(S')|$. Denoting $I = \{i_1, \dots, i_l\}$, we write $S' = S[I]$. A *factor* of S is a subword with consecutive elements of S, i.e., $s_is_{i+1}\dots s_{i+m}$, for some $1 \le i \le n$ and $0 \le m \le n - i$, we denote it $S[i, i + m]$. We sometimes use notation $[n]$ for $\{1, \dots, n\}$. We use capital letters for words, subwords, factors and sets. If S is a word over alphabet Σ and $q \in \Sigma$, we denote by $|S|_q$ the number of elements of S equal to q. The *density* $d_q(S)$ of the letter q in S is defined to be $|S|_q/|S|$.

For two subwords S' and S'' of S, we say that S' is contained in S'' if $\mathrm{supp}(S') \subseteq \mathrm{supp}(S'')$, we also denote by $S' \cap S''$ a subword of S, $S[\mathrm{supp}(S') \cap \mathrm{supp}(S'')]$. If $S = s_1 \dots s_n$ and $S[1, i] = A$, $S[i + 1, n] = B$, then we write $S = AB$ and call S a concatenation of A and B.

In all of the calculations below we omit floors and ceilings to avoid making the presentation too cluttered.

Definition 4 (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, $\varepsilon n + 1 \le i \le n - 2\varepsilon n + 1$ and every $q \in \Sigma$ it holds that*

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

Notice that in the case of the binary alphabet $\Sigma = \{0, 1\}$, $d_0(S) = 1 - d_1(S)$ and thus $|d_0(S) - d_0(S[i, i + \varepsilon n - 1])| < \varepsilon \iff |d_1(S) - d_1(S[i, i + \varepsilon n - 1])| < \varepsilon$. In this case, we shall denote $d(S) = d_1(S)$.

For example, the word

$$S = 011101010101000101001100110101010100110101010101111000010100$$

of length $n = 60$ and density $1/2$ is ε-regular for $\varepsilon = 1/5$. This could be verified directly by considering consecutive segments of length $\varepsilon n = (1/5) \cdot 60 = 12$ starting at positions $13, 14, \dots, 37$ and comparing their densities with the density $1/2$ of S. Such a 12-letter word staring at position 13 is $S' = 000101001100$, $d(S') = 8/12$, $|8/12 - 1/2| < \epsilon = 1/5$, such a 12-letter word starting at position 14 is $S'' = 001010011001$, $d(S'') = 7/12$, $|7/12 - 1/2| < 1/5$, and so on.

The notion of ε-regular words resembles the notion of pseudorandom (quasirandom) word, see [3]. However, these two notions are quite different. A word that consists of alternating 0's and 1's is ε-regular but not pseudorandom. Also, unlike the case of stronger notions of pseudorandomness, one can check in linear time, by observing at most n words $S[i, i + \varepsilon n - 1]$, whether a word is ε-regular, cf. [1] in the graph case.

Definition 5. *We call $\mathcal{S} := (S_1, \dots, S_t)$ a partition of S if $S = S_1S_2 \dots S_t$, (S is concatenation of consecutive factors S_i, $i = 1, \dots n$). A partition $\mathcal{S}$ is an ε-regular partition of a word $S \in \Sigma^n$ if*

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

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

Now we will prove and make use of a decomposition lemma, which we call Regularity Lemma for words. This is a crucial Lemma for our considerations.

Lemma 6 (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 $t_0 \le t \le T_0 = t_0 3^{1/\varepsilon^4}$.*

To prove the regularity lemma, we introduce the notion of an index and a refinement and prove a few basic facts.

Definition 7 (Index of a partition)**.** *Let $\mathcal{S} := (S_1, \dots, S_t)$ be a partition of $S \in \Sigma^n$ into consecutive factors. We define*

$$\mathrm{ind}(\mathcal{S}) = \sum_{q \in \Sigma} \sum_{i \in [t]} d_q(S_i)^2 \tfrac{|S_i|}{n}.$$

Further, for convenience we set $\mathrm{ind}_q(\mathcal{S}) = \sum_{i \in [t]} d_q(S_i)^2 \frac{|S_i|}{n}$.

Observe that $\mathrm{ind}(\mathcal{S})$ is upper-bounded by 1. Note also that an alternative way to express $\mathrm{ind}_q(\mathcal{S})$ is $\sum_{i \in [t]} \frac{|S_i|_q^2}{|S_i| n}$.

Definition 8 (Refinement of $\mathcal{S}$)**.** *Let $\mathcal{S} = (S_1, \dots, S_t)$ and $\mathcal{S}' = (S'_{1,1}, S'_{1,2}, \dots, S'_{1,m_1}, \quad S'_{2,1}, S'_{2,2}, \dots, S'_{2,m_2}, \quad \dots, \quad S'_{t,1}, S'_{t,2}, \dots, S'_{t,m_t})$ be partitions of $S \in \Sigma^n$. We say that $\mathcal{S}'$ refines $\mathcal{S}$ and write $\mathcal{S}' \preccurlyeq \mathcal{S}$, if for every $i = 1, \dots, t$, $S_i = S'_{i,1} S'_{i,2} \cdots S'_{i,m_i}$.*

Claim 9. *Let $\mathcal{S}$ and $\mathcal{S}'$ be partitions of $S \in \Sigma^n$. If $\mathcal{S}' \preccurlyeq \mathcal{S}$ then*

$$\mathrm{ind}(\mathcal{S}') \ge \mathrm{ind}(\mathcal{S}).$$

Proof. Let $\mathcal{S} = (S_1, \dots, S_t)$ and

$$\mathcal{S}' = (S'_{1,1}, S'_{1,2}, \dots, S'_{1,m_1}, \quad S'_{2,1}, S'_{2,2}, \dots, S'_{2,m_2}, \quad \dots, \quad S'_{t,1}, S'_{t,2}, \dots, S'_{t,m_t}).$$

We proceed for each $q \in \Sigma$ as follows:

$$\begin{aligned}
\mathrm{ind}_q(\mathcal{S}') &= \sum_{S' \in \mathcal{S}'} d_q(S')^2 \frac{|S'|}{n} \\
&= \sum_{i=1}^{t} \sum_{j=1}^{m_i} d_q(S'_{i,j})^2 \frac{|S'_{i,j}|}{n} \\
&= \sum_{i=1}^{t} \frac{|S_i|}{n} \sum_{j=1}^{m_i} d_q(S_{i,j})^2 \frac{|S'_{i,j}|}{|S_i|} \\
&\overset{\text{Jensen's inequality}}{\ge} \sum_{i=1}^{t} \frac{|S_i|}{n} \left(\sum_{j=1}^{m_i} d_q(S'_{i,j}) \frac{|S'_{i,j}|}{|S_i|} \right)^2 \\
&= \sum_{i=1}^{t} \frac{|S_i|}{n} \left(\sum_{j=1}^{m_i} \frac{|S'_{i,j}|_q}{|S'_{i,j}|} \frac{|S_{i,j}|}{|S_i|} \right)^2 \\
&= \sum_{i=1}^{t} \frac{|S_i|}{n} d_q(S_i)^2 \\
&= \mathrm{ind}_q(\mathcal{S}).
\end{aligned}$$

Now, building the sum over all $q \in \Sigma$ yields:

$$\operatorname{ind}(\mathcal{S}') \geq \operatorname{ind}(\mathcal{S}).$$

□

The next claim shows that if a word S is not ε-regular, then there is a refinement of (S) whose index exceeds the index of (S) by at least ε^3.

Claim 10. *Let $S \in \Sigma^m$ be an ε-irregular word. Then there is a partition (A, B, C) of S such that $|A|, |B|, |C| \geq \varepsilon m$ and*

$$\operatorname{ind}((A,B,C)) \geq \operatorname{ind}((S)) + \varepsilon^3 = \left(\sum_{q\in\Sigma} d_q(S)^2\right) + \varepsilon^3. \tag{2}$$

Proof. Since S is not ε-regular, there exists an element $q \in \Sigma$ and an i with $\varepsilon m+1 \leq i \leq m - 2\varepsilon m + 1$ such that $|d - d(S[i, i + \varepsilon m - 1])| \geq \varepsilon$, where $d := d_q(S)$ and $d(T) := d_q(T)$ for any factor T of S. Assume first that $d - d(S[i, i + \varepsilon m - 1]) \geq \varepsilon$ and set $\gamma := d - d(S[i, i + \varepsilon m - 1])$, $A := S[1, i - 1]$, $B := S[i, i + \varepsilon m - 1]$ and $C := S[i + \varepsilon m, m]$, $a := |A|$, $b := |B| = \varepsilon m$ and $c := |C|$.

Observe further that

$$|S|_q = d(A)a + d(B)b + d(C)c = dm, \quad d((A,C)) = \tfrac{dm-(d-\gamma)b}{a+c}, \quad d(B) = d - \gamma.$$

Since $a + c = m - b$ and $\operatorname{ind}_q((A,B,C)) = \operatorname{ind}_q((A,C,B))$,

$$\begin{aligned}
\operatorname{ind}_q((A,B,C)) &\geq d((A,C))^2 \frac{a+c}{m} + d(B)^2 \frac{b}{m} \\
&= \left(\frac{dm - (d-\gamma)b}{a+c}\right)^2 \frac{a+c}{m} + (d-\gamma)^2 \frac{b}{m} \\
&= \frac{(dm-(d-\gamma)b)^2}{(m-b)m} + (d-\gamma)^2 \frac{b}{m} \\
&= \frac{1}{(m-b)m}\left[d^2(m^2 - mb) + \gamma^2(mb)\right] \\
&= d^2 + \frac{\gamma^2 b}{m-b} \geq d^2 + \frac{\varepsilon^3 m}{(1-\varepsilon)m} \geq d^2 + \varepsilon^3.
\end{aligned}$$

The case when $d - d(S[i, i + \varepsilon n - 1]) \leq -\varepsilon$ works out similarly. Indeed, set $\gamma := d - d(S[i, i + \varepsilon m - 1])$ as before and notice that $|\gamma| \geq \varepsilon$ and all the computations above are exactly the same.

So, $\operatorname{ind}_q((A,B,C)) \geq d_q^2 + \varepsilon^3$. For all other $q' \in \Sigma$, Claim 9 gives that $\operatorname{ind}_{q'}((A,B,C)) \geq \operatorname{ind}_{q'}((S)) = d_{q'}^2(S)$. Thus

$$\operatorname{ind}((A,B,C)) = \operatorname{ind}_q((A,B,C)) + \sum_{q'\in\Sigma-\{q\}} \operatorname{ind}_{q'}((A,B,C)) \geq \sum_{q'\in\Sigma} d_{q'}(S)^2 + \varepsilon^3.$$

□

Finally we are in position to finish the argument.

Proof of the Regularity Lemma for Words. Take $\varepsilon > 0$ and t_0 as given. Let $n_0 = t_0 \varepsilon^{-\varepsilon^{-4}}$ and $n > n_0$. Suppose that we have a word $S \in \Sigma^n$. Split it into t_0 consecutive factors $S_1, \ldots, S_{t_0}$ of the same length $\frac{n}{t_0}$. If $\mathcal{S} := (S_1, \ldots, S_{t_0})$ is not

an ε-regular partition, then let $I \subseteq [t_0]$ be the set of all indices such that, for every $i \in I$, S_i is not ε-regular (thus, $\sum_{i\in I} |S_i| \geq \varepsilon n$). Then, by Claim 10 we can refine each S_i, $i \in I$, into factors A_i, B_i and C_i such that $|A_i|, |B_i|, |C_i| \geq \varepsilon |S_i|$ and $\mathrm{ind}((A_i, B_i, C_i)) \geq \sum_{q\in\Sigma} d_q(S_i)^2 + \varepsilon^3$ (in the case that (1) is violated for several $q \in \Sigma$, choose an arbitrary such q). We perform such refinement for each S_i, $i \in I$, obtaining a partition $\mathcal{S}' \preccurlyeq \mathcal{S}$, noticing that

$$\begin{aligned}
\mathrm{ind}(\mathcal{S}') &= \sum_{q\in\Sigma} \sum_{j\in[t_0]\setminus I} d_q(S_j)^2 \frac{|S_j|}{n} + \\
&\qquad \sum_{q\in\Sigma} \sum_{i\in I} \left(d_q(A_i)^2 \frac{|A_i|}{n} + d_q(B_i)^2 \frac{|B_i|}{n} + d_q(C_i)^2 \frac{|C_i|}{n} \right) \\
&= \sum_{q\in\Sigma} \sum_{j\in[t_0]\setminus I} d_q(S_j)^2 \frac{|S_j|}{n} + \sum_{i\in I} \mathrm{ind}((A_i, B_i, C_i)) \frac{|S_i|}{n} \\
&\overset{(2)}{\geq} \sum_{q\in\Sigma} \sum_{j\in[t_0]\setminus I} d_q(S_j)^2 \frac{|S_j|}{n} + \sum_{i\in I} (\mathrm{ind}((S)) + \varepsilon^3) \frac{|S_i|}{n} \\
&= \mathrm{ind}(\mathcal{S}) + \varepsilon^3 \frac{\sum_{i\in I} |S_i|}{n} \\
&\geq \mathrm{ind}(\mathcal{S}) + \varepsilon^4.
\end{aligned}$$

Thus, $\mathcal{S}'$ refines $\mathcal{S}$ and has higher index. If $\mathcal{S}'$ is not an ε-regular partition of S, then we can repeat the procedure above by refining $\mathcal{S}'$ etc. Recall that an index of any partition $\mathcal{S}$ is upper-bounded by 1. Since the index is increased by at least ε^4 at each iteration, we have to stop either when the factors or the partition are too small, or when the index reaches 1.

Let η be the number of iterations, $\eta \leq 1/\varepsilon^4 = \varepsilon^{-4}$. The length of each factor in the partition decreases at a rate of at most ε at each iteration. So, after η iterations, the length of each factor is at least

$$\frac{n}{t_0}\varepsilon^\eta \geq \frac{n_0}{t_0}\varepsilon^\eta \geq \frac{t_0 \varepsilon^{-\varepsilon^{-4}}}{t_0} \varepsilon^{\varepsilon^{-4}} = 1.$$

Notice that such a partition consists of at most $3^\eta t_0 \leq 3^{1/\varepsilon^4} t_0$ words, since at each iteration each of the words is partitioned into at most 3 new ones. Therefore, $T_0 \leq 3^{1/\varepsilon^4} t_0$ and each factor in the partition has length at least $t_0^{-1} \varepsilon^{1/\varepsilon^4} n$. □

Remark. Improved bounds for the Regularity Lemma are given in Section 6.2. We do not provide these arguments in the proof of the lemma here for readability reasons.

3. Proof of Theorem 1.

Before we prove our main theorem about binary words, we show a useful claim about twins in ε-regular words.

Claim 11. *If S is an ε-regular word, then $2f(S) \geq |S| - 5\varepsilon|S|$.*

Proof. Let $|S| = m$. We partition S into $t = 1/\varepsilon$ consecutive factors $S_1, \ldots, S_{1/\varepsilon}$, each of length εm. Since S is ε-regular, $|d(S_i) - d(S)| < \varepsilon$, for every $i \in \{2, \ldots, 1/\varepsilon - 1\}$. Thus each S_i has at least $(d(S) - \varepsilon)\varepsilon m$ occurrences of 1's and at least $(1 - d(S) - \varepsilon)\varepsilon m$ occurrences of 0's. Let $S_i(1)$ be a subword of S_i consisting of exactly

$(d(S)-\varepsilon)\varepsilon m$ letters 1 and $S_i(0)$ be a subword of S_i consisting of exactly $(1-d(S)-\varepsilon)\varepsilon m$ letters 0. When t is even, consider the following two disjoint subwords of S: $A = S_2(1)S_3(0)S_4(1)\cdots S_{t-2}(1)$ and $B = S_3(1)S_4(0)S_5(1)\cdots S_{t-2}(0)S_{t-1}(1)$. When t is odd, A and B are constructed similarly.

We see that A and B together have at least $m - 2\varepsilon^2 m(1/\varepsilon - 3) - 3\varepsilon m$ elements, where $2\varepsilon^2 m(1/\varepsilon - 3)$ is an upper bound on the number of 0's and 1's which we had to "throw away" to obtain *exactly* $(d(S)-\varepsilon)\varepsilon m$ letters 1 and $(1-d(S)-\varepsilon)\varepsilon m$ letters 0 in each S_i, $2\varepsilon m$ is the number of elements in S_1 and S_t, and εm is the upper bound on $|S_2(0)| + |S_{t-1}(1)|$.

Thus, $2f(S) \geq m - 5\varepsilon m$. This concludes the proof of the claim. □

Notice that we could slightly improve on $5\varepsilon m$ above by finding twins of size $\varepsilon m/3$ each in S_1 and S_t in an already mentioned way.

Proof of Theorem 1. For the lower bound, consider S, a binary word of length n. Let $\varepsilon = C(\frac{\log n}{\log\log n})^{-1/4}$ and $t_0 := \frac{1}{\varepsilon}$. We see that $n \geq \varepsilon^{-\varepsilon^{-4}}$. Applying the Regularity Lemma for Words gives an ε-regular partition $S = S_1, \ldots, S_t$ with $1/\varepsilon \leq t \leq T_0 = t_0 3^{1/\varepsilon^4}$. We apply Claim 11 to every ε-regular factor S_i. Furthermore, since S_i's appear consecutively in S, we can put the twins from each of S_i's together obtaining twins for the whole word S. This way we see:

$$2f(S) \geq \sum_{\substack{i\in[t]\\ S_i \text{ is } \varepsilon-\text{regular}}} (|S_i| - 5\varepsilon|S_i|) \geq n - 5\varepsilon n - \varepsilon n = n - 6\varepsilon n,$$

here εn corresponds to the total lengths of not ε-regular factors.

Next we shall prove the upper bound on $f(n,\{0,1\})$ by constructing a binary word S such that $2f(S) \leq |S| - \log|S|$. Let $S = S_k S_{k-1}\ldots S_0$, where $|S_i| = 3^i$, S_i consists only of 1's for even i, and it consists only of 0's for odd i's. I.e., S is built of iterated 1- or 0-blocks exponentially decreasing in size. Let A and B be twins in S.

Assume first that A and B have the same number of elements in S_k. Since S_k has odd number of elements, and A, B restricted to $S' = S_{k-1}S_{k-2}\cdots S_0$ are twins, by induction we have that $|A|+|B| \leq (|S_k|-1)+(|S'|-\log(|S'|)) = |S|-1-\log(|S'|) \leq |S| - \log|S|$. That is true since $|S_k| = 3^k$, $|S| = (3^{k+1}-1)/2$.

Now assume, w.l.o.g. that A has more elements in S_k than B in S_k. Then B has no element in S_{k-1}. We have that $|A\cap S_{k-1}| \geq |S_{k-1}|/2$, otherwise $|A|+|B| \leq |S| - |S_{k-1}|/2 \leq |S| - \log|S|$. So, $s = |A\cap S_{k-1}| \geq |S_{k-1}|/2 \geq 3^{k-1}/2$, and s elements of B must be in $S_{k-3}\cup S_{k-5}\cdots$. But $|S_{k-3}| + |S_{k-5}| + \cdots \leq 3^{k-2}/2$, a contradiction. □

Remark. One can find words of length $n/2 - o(n)$ as follows from the proof of Theorem 1 by an algorithm with $O(\varepsilon^{-4}n) = O(\frac{n\log n}{\log\log n})$ steps.

4. k-twins over alphabet of at most k letters

Proof of Theorem 2. t As before, we concentrate first on ε-regular words. Let S be an ε-regular word of length n over alphabet $\Sigma = \{0,\ldots,\ell-1\}$ and recall the

assumption $\ell \le k$. We partition S in $t = 1/\varepsilon$ consecutive factors $S_1, \dots, S_{1/\varepsilon}$, each of length εn. Since S is ε-regular, $|d_q(S_i) - d_q(S)| < \varepsilon$, for every $i \in \{2, \dots, 1/\varepsilon - 1\}$, and every $q \in \Sigma$. Thus S_i has at least $(d_q(S) - \varepsilon)\varepsilon n$ letters q, for each $q \in \Sigma$.

We construct k-twins $A_1, \dots, A_k$ as follows. Each of A_j's consists of consecutive blocks, with first block consisting of $(d_0(S) - \varepsilon)\varepsilon n$ letters 0, followed by a block of $(d_1(S) - \varepsilon)\varepsilon n$ letters 1, ..., followed by a block of $(d_{\ell-1}(S) - \varepsilon)\varepsilon n$ letters $\ell - 1$, followed by a block of $(d_0(S) - \varepsilon)\varepsilon n$ letters 0, and so on.

Since $k \ge |\Sigma|$, we will use all but at most $\frac{1}{\varepsilon}\varepsilon^2 n|\Sigma| + (2|\Sigma|)\varepsilon n = 3|\Sigma|\varepsilon n$ elements, where the first summand accounts for the number of elements that we did not use when choosing exactly $(d_q(S) - \varepsilon)\varepsilon n$ elements q from each S_i and each $q \in \Sigma$ and the second summand for the number of elements in $S_1, \dots, S_\ell$, and from $S_{1/\varepsilon - \ell + 1}, \dots, S_{1/\varepsilon}$.

Below are the examples in the special cases when $|\Sigma| = \ell = k$ and when $|\Sigma| = 2$ and $k = 4$.

Example 1.

$$
\begin{aligned}
A_1 &= S_2(0)S_3(1)S_4(2)\cdots S_{\ell+1}(\ell-1)S_{\ell+2}(0)S_{\ell+3}(1)\cdots S_{2\ell+1}(\ell-1)\cdots,\\
A_2 &= \quad S_3(0)S_4(1)S_5(2)\cdots S_{\ell+2}(\ell-1)S_{\ell+3}(0)S_{\ell+4}(1)\cdots S_{2\ell+2}(\ell-1)\cdots,\\
&\vdots\\
A_i &= \quad\quad S_{i+1}(0)S_{i+2}(1)S_{i+3}(2)\cdots S_{i+\ell}(\ell-1)S_{i+\ell+1}(0)S_{i+\ell+2}(1)\cdots S_{i+2\ell}(\ell-1)\cdots\\
&\vdots\\
A_k &= \quad\quad\quad\quad\quad\quad S_{\ell+1}(0)S_{\ell+2}(1)S_{\ell+3}(2)\cdots S_{2\ell}(\ell-1)S_{2\ell+1}(0)\cdots S_{3\ell}(\ell-1)\cdots
\end{aligned}
$$

Example 2.

$$
\begin{aligned}
A_1 &= S_2(0)S_3(1) \qquad\quad S_6(0)S_7(1)\cdots\\
A_2 &= \quad S_3(0)S_4(1) \qquad\quad S_7(0)S_8(1)\cdots\\
A_3 &= \quad\quad S_4(0)S_5(1) \qquad\quad S_8(0)S_9(1)\cdots\\
A_4 &= \quad\quad\quad S_5(0)S_6(1) \qquad\quad S_9(0)S_{10}(1)\cdots
\end{aligned}
$$

Here $S_i(j)$ is the block of $(d_j(S) - \varepsilon)\varepsilon n$ letters j taken from S_i. So, in general, the total number of elements in $A_1, \dots, A_k$ is at least $n - 3|\Sigma|\varepsilon n$.

Thus, $kf(S) \ge n - 3|\Sigma|\varepsilon n$.

To provide the lower bound on $f(n, k, \Sigma)$ we proceed as in the proof of Theorem 1 by first finding a regular partition of a given word and then applying the above construction to regular factors with an appropriate choice of ε. Construction for the upper bound is similar to one from Theorem 1. □

5. Large alphabets and small k-twins

Proof of Theorem 3. The proof of the lower bound proceeds by considering a scattered subword W consisting of the k most frequent letters. Clearly, $|W| \ge \frac{k}{|\Sigma|}n$, which together with Theorem 2 yields the lower bound.

The upper bound we obtain is immediate from Theorem 1 and from computing the expected number of k-twins of length m each in a random word of length n over an alphabet Σ, denote $|\Sigma| = \ell$. If the expectation if less than 1, this means

that there is a word S with $f(S,k) < m$. Indeed, there are

$$\frac{1}{k!}\prod_{i=0}^{k-1}\binom{n-im}{m}$$

distinct sets of k disjoint subwords each of length m in a word of length n. The probability that such a set corresponds to k-twins, when each letter is chosen with probability $1/\ell$ independently, is $\ell^{(1-k)m}$. Thus, the expected number of k-twins is at most

$$\ell^{(1-k)m}\prod_{i=0}^{k-1}\binom{n-im}{m} = \ell^{-(k-1)m}\frac{n!}{(m!)^k(n-km)!} \le \ell^{-(k-1)m}\frac{n^n}{m^{km}(n-km)^{n-km}},$$

that is, for $m=\alpha n$, is at most

$$\ell^{-(k-1)\alpha n}\frac{n^n}{(\alpha n)^{k\alpha n}(n-k\alpha n)^{n-k\alpha n}} = \left(\ell^{-(k-1)\alpha}\alpha^{-k\alpha}(1-k\alpha)^{k\alpha-1}\right)^n.$$

Thus, if $\ell^{-(k-1)\alpha}\alpha^{-k\alpha}(1-k\alpha)^{k\alpha-1}$ is less than 1 then $f(S,k)\le \alpha n$.

In particular, for $k=2$ and $\ell=5$ one can compute that $\alpha < 0.49$. Further, for $k=2$, it follows by simple estimates that $\alpha = O(1/\sqrt{\ell})$.

□

6. Concluding Remarks

6.1. **Small values of $f(n,k,\Sigma)$.** We will slightly abuse notation and denote by $f(n,k,\ell)$ the value of $f(n,k,\Sigma)$ with $|\Sigma|=\ell$. In the introductory section it was observed that $f(3,2,2)=1$ yielding immediately a weak lower bound on $f(n,2,2)$ to be $\lfloor n/3\rfloor$. In general, it holds clearly that

$$f(n,k,\ell) \ge \left\lfloor \tfrac{n}{m}\right\rfloor f(m,k,\ell).$$

For example, we determined (Theorem 3) a lower bound on $f(n,2,3)$ to be $\frac{1}{3}n-o(n)$. We do not know whether it is tight and, moreover, whether one can achieve it, without $o(n)$ term, by finding a (reasonable) number t such that $f(t,2,3)\ge \frac{t}{3}$. If one could find such t this would immediately give another proof of $f(n,2,3)\ge \frac{1}{3}n-t$. However, the smallest value for such possible t could be 21, which already presents a computationally challenging task. In the tables above we summarize estimates on the values on $f(n,k,\ell)$, which were determined with the help of a computer. We see that for $n=21$, $f(n,2,3)\le 7 = n/3$, for $n=22$, $f(n,2,3)\le 7 < n/3$. Thus, the first "open" case which might improve lower bound $n/3$ on $f(n,2,3)$ is $f(23,2,3)$.

$\Sigma\backslash n$	6	7	8	9	10	11	12	13	14	15	16	17
$\{0,1\}$	2	2	2	3	3	4	4	5	5	5	6	6
$\{0,1,2\}$	1	1	2	2	2	3	3	3	4	4	4	4

$\Sigma\backslash n$	18	19	20	21	22	23	24
$\{0,1\}$	7	7	8				
$\{0,1,2\}$	≤ 5	≤ 6	≤ 6	≤ 7	≤ 7	≤ 8	≤ 8

Table 1. Values for small t of $f(t,2,2)$ and $f(t,2,3)$.

6.2. **Improving the $O\left(|\Sigma|\left(\frac{\log\log n}{\log n}\right)^{1/4}\right)n$ term.** Further we remark, that a more careful analysis below of the increment argument in the proof of Lemma 6 leads to the bound $T_0 \le t_0 3^{(-2\log\varepsilon)/\varepsilon^3}$, which in turn improves the bounds in Theorems 1 and 2 to

$$\left(1 - C|\Sigma|\left(\frac{(\log\log n)^2}{\log n}\right)^{1/3}\right)n \le kf(n,k,\Sigma).$$

Recall that in the proof of Lemma 6 we set up an index and refining a corresponding partition each time we increase it by at least ε^4. Let's reconsider jth refinement step at which the partition $\mathcal{S} = (S_1, \dots, S_{t_0})$ is to be refined. Further recall that I consists of the indices i such that S_i is not ε-regular. Let α_j be such that

$$\sum_{i\in I} |S_i| = \alpha_j n. \tag{3}$$

In the original proof we iterate as long as $\alpha_j \ge \varepsilon$ holds. And by peforming an iteration step we merely use the fact that $\alpha_j \ge \varepsilon$ which leads to ε^4 increase of the index during one iteration step. Recall that $\mathrm{ind}(\mathcal{S})$ was defined as follows:

$$\mathrm{ind}(\mathcal{S}) = \sum_{q\in\Sigma}\sum_{j\in[|\mathcal{S}|]} d_q(S_j)^2\frac{|S_j|}{n},$$

and for each further refinement $\mathcal{S}' \preccurlyeq \mathcal{S}$ it holds:

$$\mathrm{ind}(\mathcal{S}) \le \mathrm{ind}(\mathcal{S}') = \frac{(1-\alpha_j)n}{n}\mathrm{ind}(\mathcal{S}_1) + \frac{\alpha_j n}{n}\mathrm{ind}(\mathcal{S}_2) \le \sum_{q\in\Sigma}\sum_{j\in[|\mathcal{S}|]\setminus I} d_q(S_j)^2\frac{|S_j|}{n} + \alpha_j, \tag{4}$$

where $\mathcal{S}_1$ consists of ε-regular words from $\mathcal{S}$(these words are not partitioned/refined anymore) and $\mathcal{S}_2$ consists of not ε-regular words from $\mathcal{S}$ (and their lengths sum up to $\alpha_j n$).

Let m be the total number of iteration steps until we arrive at an ε-regular partition. Let $\alpha_1, \dots, \alpha_m$ be the numbers, where $\alpha_j n$ is the sum over the lengths of not ε-regular words in the partition at step j, $j \in [m]$ (cf.(3)).

By the discussion above

$$1 \ge \alpha_1 \ge \alpha_2 \ge \dots \ge \alpha_m \ge \varepsilon.$$

Next, we partition $(\varepsilon, 1]$ into $\log_2\frac{1}{\varepsilon}$ consecutive intervals $(y_{i+1}, y_i]$ where $y_1 = 1$ and $y_{i+1} = y_i/2$. We claim that each interval $(y_{i+1}, y_i]$ contains at most $\frac{2}{\varepsilon^3}$ α_js. Indeed, the increase of the index during step j where $\alpha_j \in (y_{i+1}, y_i]$ is at least

$$\alpha_j\varepsilon^3 > y_{i+1}\varepsilon^3.$$

Further, let j' be the smallest index such that $\alpha_{j'} \le y_i$ and j'' be the largest index such that $\alpha_{j''} > y_{i+1}$. Let ind_j be the index before the jth refinement step. Then by (4) the following holds for $j' + 1 \le j \le j''$:

$$\mathrm{ind}_{j'+1} \le \mathrm{ind}_j \le \mathrm{ind}_{j''} \le \mathrm{ind}_{j'+1} + y_i.$$

This implies that the number of α_js in the interval $(y_{i+1}, y_i]$ cannot be bigger than

$$\frac{y_i}{y_{i+1}\varepsilon^3} = \frac{2}{\varepsilon^3}.$$

Thus, we obtain the following upper bound on m

$$m \leq \frac{2\log_2 \frac{1}{\varepsilon}}{\varepsilon^3},$$

which leads to $T_0 \leq t_0 3^{(-2\log\varepsilon)/\varepsilon^3}$, $n_0 = t_0\varepsilon^{-(2\log 1/\varepsilon)/\varepsilon^3}$ and thus we can regularize with $\varepsilon = \left(\frac{(\log\log n)^2}{\log n}\right)^{1/3}$.

6.3. **Conclusions.** In the paper we studied the length of largest twins in words. One of the most interesting questions is whether any given word of length n over alphabet Σ has k-twins of size $n(1-o(1))/k$ each, i.e., when the k-twins cover almost all letters of the word. In this case, we say that the pair $(k, |\Sigma|)$ is *good*. We have shown that $(k, |\Sigma|)$ is good for $k \geq |\Sigma|$. Also, we observed that some pairs $(k, |\Sigma|)$ are not good for $k < |\Sigma|$. The smallest such pair of values for $|\Sigma|$ and k that we know is $(k, |\Sigma|) = (2, 5)$. One of the most fascinating open questions is to determine whether the pairs $(k, |\Sigma|) = (k, k+1)$ are good, even in the case $k = 2$.

Acknowledgements

The authors would like to thank Sergey Avgustinovich and Boris Bukh for fruitful discussions and several nice observations and the referees for careful reading and comments improving the presentation of the results.

References

[1] N. Alon, R. A. Duke, H. Lefmann, V. Rödl, and R. Yuster, The algorithmic aspects of the regularity lemma, J. Algorithms 16 (1994), no. 1, 80–109. 2
[2] C. Choffrut, J. Karhumäki, Combinatorics of words. In: Handbook of Formal Languages, Springer, 1997.
[3] F. R. K. Chung, R. L. Graham, Quasi-random subsets of $\mathbb{Z}_n$, J. Combin. Theory Ser. A 61 (1992), no. 1, 64–86. 2
[4] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, MIT Press and McGraw-Hill (2001), 350–355. 1
[5] M. Dudík, L. J. Schulman: Reconstruction from subsequences. J. Comb. Theory, Ser. A 103(2) (2003), 337–348. 1
[6] W. T. Gowers, A new proof of Szemerédi's theorem, Geom. Funct. Anal. 11 (2001), no. 3, 465–588. 1
[7] D. S. Hirschberg, A linear space algorithm for computing maximal common subsequences, Communications of the ACM 18 (6) (1975), 341–343. 1
[8] J. Komlós, M. Simonovits, Szemerédi's regularity lemma and its applications in graph theory. In: Combinatorics, Paul Erdős is Eighty, Vol. 2 (Keszthely, 1993), volume 2 of Bolyai Soc. Math. Stud., pp. 295352. János Bolyai Math. Soc., Budapest, 1996. 1
[9] M. Lothaire, Algebraic combinatorics on words. Cambridge University Press, 2002.
[10] A. Mateescu, A. Salomaa, and S. Yu. Subword histories and parikh matrices. J. Comput. Syst. Sci., 68(1):1–21, 2004. 1
[11] A. Salomaa, Counting (scattered) subwords. Bulletin of the EATCS, 81:165–179, 2003. 1
[12] E. Szemerédi, *Regular partitions of graphs.* (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), pp. 399–401, Colloq. Internat. CNRS, 260, CNRS, Paris, 1978. 1
[13] X. Xia, Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics. New York: Springer, 2007. 1

Iowa State University, Ames, U.S.A. and Karlsruher Institut für Technologie, Karlsruhe, Germany
E-mail address: maria.aksenovich@kit.edu

Freie Universität Berlin, Institut für Mathematik, Berlin, Germany
E-mail address: person@math.fu-berlin.de

University of Turku, Turku, Finland, and Sobolev Institute of Mathematics, Novosibirsk, Russia
E-mail address: svepuz@utu.fi