Every binary word is, almost, a shuffle of twin subsequences — a theorem of Axenovich, Person and Puzynina

Martin Klazar*

August 17, 2015

A *twin* in a word $u = a_1a_2 \ldots a_n$ is a pair (u_1, u_2) of disjoint and identical ($u_1 = u_2$) subsequences of u. A *binary word* is a word $u \in \{0, 1\}^n$. For example, $\underline{1}0\overline{10}\underline{0}$ is a binary word with a twin $(u_1, u_2) = (a_1a_5, a_3a_4)$ (marked by under(over)linings). By $|u|$ we denote the length of a (binary) word u. We shall consider only binary words here and will often omit the adjective.

Theorem (Axenovich, Person and Puzynina, 2013). *For every $\varepsilon > 0$ there is an n_0 such that every binary word of length $n > n_0$ has a twin (u_1, u_2) with $n - 2|u_1| < \varepsilon n$.*

That is, for given $\varepsilon > 0$ every sufficiently long word u has a partition $u = u_1 \cup u_2 \cup u_3$ into three subsequences such that $u_1 = u_2$ and $|u_3| < \varepsilon|u|$. Here is a simple argument giving $\varepsilon \doteq 1/3$. We partition u into intervals (factors) I_i of length 3 each and a residual interval J with $|J| \le 2$. Each I_i contains two 0s or two 1s. For each I_i we put one of them in u_1 and the other in u_2 and set u_3 to be the rest of u. Then $u_1 = u_2$ and $|u_3| = \lfloor |u|/3 \rfloor + |J| < |u|/3 + 2$. Can you decrease the 1/3?

The purpose of this text, written in very hot Prague days, is to enjoy and advertise the beautiful result of Axenovich, Person and Puzynina [2] lying on the border of combinatorics on words and Ramsey theory, and possibly include it later in the prepared book [3]. The theorem is remarkable for its beauty and simplicity, it is indeed a bit surprising that it was discovered only recently, and for the fact that its proof uses a particularly technically simple and clear version of the regularity lemma method, simpler than the usual graph-theoretical setting, not speaking of hypergraph versions. We write more on [2] and the proof at the end.

A regularity lemma for binary words

For $\varepsilon \in (0, 1)$ and $u = a_1a_2 \ldots a_n$ a binary word, an ε-*interval in* u is an interval $I = a_ia_{i+1} \ldots a_{i+j}$ of length $|I| = j + 1 = \lceil \varepsilon n \rceil$. For $i \in \{0, 1\}$ we

*klazar@kam.mff.cuni.cz

define $d_i(u) = \#(j, a_j = i)/|u| \in [0, 1]$, the density of the letter i in u. We set $d(u) = d_1(u)$. Clearly, $d_1(u) + d_0(u) = 1$ and this makes binary words simpler than words over larger alphabets as it suffices to work with just one density $d(u) = d_1(u)$.

Definition (ε-regularity). *Let $\varepsilon \in (0, 1)$ and $u \in \{0, 1\}^n$. We say that u is ε-regular if*

$$|d(u) - d(I)| < \varepsilon$$

whenever I is ε-interval in u. A partition $u = u_1u_2 \dots u_t$ into intervals is ε-regular if the total length of non-ε-regular intervals is small,

$$\sum_{u_i \text{ is not } \varepsilon\text{-reg.}} |u_i| < \varepsilon n = \varepsilon |u| \ .$$

In an interval partition $u = u_1u_2 \dots u_t$, as in the above definition, all u_i are nonempty, if it is not said else. Note that the definition of ε-regularity of a word is equivalent to one with $d(\cdot)$ replaced by $d_i(\cdot)$, $i = 0, 1$. As an example note that the partition of u into singleton intervals is always ε-regular as each singleton word is ε-regular.

Let us show that ε-regular words have large twins.

Lemma 1. *Let $\varepsilon \in (0, 1)$. Every ε-regular binary word u has a twin (u_1, u_2) with $|u| - 2|u_1| < 5\varepsilon|u| + 3$.*

Proof. We set $m = \lceil \varepsilon |u| \rceil$, $d_i = \lceil (d_i(u) - \varepsilon)m \rceil$ for $i = 0, 1$, and partition u as $u = v_1v_2 \dots v_t$ so that $|v_i| = m$ for $i < t$ and v_t may be empty with $|v_t| < m$. For each $i = 1, 2, \dots, t-1$, by ε-regularity v_i contains $\geq d_1$ ones and $\geq d_0$ zeros. We put in u_1 some d_1 ones from v_1, some d_0 zeros from v_2, some d_1 ones from v_3, some d_0 zeros from v_4 and so on in the alternating fashion up to v_{t-2}. We define the other twin u_2 in much the same way, but use intervals $v_2, v_3, \dots, v_{t-1}$. (If $d_i < 0$, no problem, we replace it by 0.) The rest of u ends in the bin, the subsequence u_3. By construction, u_1 and u_2 form a twin in u, are disjoint and identical subsequences. Since $d_0 + d_1 \geq m(1 - 2\varepsilon)$ and $tm < |u|$,

$$|u_3| \leq |v_1| + |v_{t-1}| + |v_t| + (t - 3)2\varepsilon m < 3m + 2\varepsilon|u| < 5\varepsilon|u| + 3 \ .$$

□

The *index* $\mathrm{ind}(P)$ of an interval partition P of a word u given by $u = u_1u_2 \dots u_t$ is

$$\mathrm{ind}(P) = \frac{1}{|u|} \sum_{i=1}^{t} d(u_i)^2 |u_i| \ .$$

Since $\sum_{i=1}^{t} \frac{|u_i|}{|u|} = 1$ and $\frac{|u_i|}{|u|}, d(u_i) \in [0, 1]$, $\mathrm{ind}(P) \in [0, 1]$ as well.

Lemma 2 (regularity lemma). *Let $\varepsilon \in (0,1)$, $t_0 \geq 1$ be an integer, and $T_0 = t_0 3^{1/\varepsilon^4}$. Then every word $u \in \{0,1\}^n$ with $n \geq t_0$ has an ε-regular partition into t intervals with $t_0 \leq t \leq T_0$.*

Proof. First we show that index does not decrease when a partition is refined (we need actually only very particular case of this inequality), and then how to increase index by splitting a non-ε-regular word in two or three intervals. Finally we obtain the desired ε-regular partition by iterating the splitting.

For the first part, let $u = u_1 u_2 \ldots u_t = \ldots u_{i,j} \ldots$, $1 \leq i \leq t$ and $1 \leq j \leq t_i$, be an interval partition P of a binary word into t intervals and its refinement R given by t interval partitions $u_i = u_{i,1} u_{i,2} \ldots u_{i,t_i}$. We show that $\mathrm{ind}(R) \geq \mathrm{ind}(P)$. Indeed,

$$\mathrm{ind}(R) = \sum_{i=1}^{t} \frac{|u_i|}{|u|} \sum_{j=1}^{t_i} \frac{d(u_{i,j})^2 |u_{i,j}|}{|u_i|} \geq \sum_{i=1}^{t} \frac{|u_i|}{|u|} \left(\sum_{j=1}^{t_i} \frac{d(u_{i,j}) |u_{i,j}|}{|u_i|} \right)^2 = \mathrm{ind}(P)$$

by Jensen's inequality applied to $f(x) = x^2$ and since the last inner sum equals $d(u_i)$ (by the definition of density).

For a non-ε-regular binary word u we find an interval partition $u = u_1 u_2 u_3$ such that u_1 or u_3 but not both may be empty and

$$\mathrm{ind}(u_1 u_2 u_3) \geq \mathrm{ind}(u) + \varepsilon^3 = d(u)^2 + \varepsilon^3 \ .$$

We define it by setting u_2 to be an ε-interval in u such that $|d(u) - d(u_2)| \geq \varepsilon$. We denote $d = d(u)$, $\gamma = d - d(u_2)$, so $|\gamma| \geq \varepsilon$, $m = |u|$, $a = |u_1|$, $b = |u_2| = \lceil \varepsilon m \rceil$, and $c = |u_3|$. Then, by part 1, $\mathrm{ind}(u_1 u_2 u_3)$ is at least

$$d(u_1 u_3)^2 \frac{a+c}{m} + d(u_2)^2 \frac{b}{m} = \left(\frac{dm - (d-\gamma)b}{a+c} \right)^2 \frac{a+c}{m} + (d-\gamma)^2 \frac{b}{m}$$

which after replacing $a + c = m - b$ simplifies to

$$d^2 + \frac{\gamma^2 b}{m-b} \geq d^2 + \frac{\varepsilon^3 m}{m} = d^2 + \varepsilon^3 \ .$$

Finally, let ε, t_0, n, and u be as given. We start with any partition S of u into t_0 intervals u_i. Let I be the indices i with u_i not ε-regular. If S is not ε-regular ($\sum_{i \in I} |u_i| \geq \varepsilon |u|$) we split each u_i with $i \in I$ into $u_i = u_{i,1} u_{i,2} u_{i,3}$ as described in part 2. The resulting interval partition P of u satisfies by part 2 (for $u_{i,j}=\emptyset$ we set $d(u_{i,j})=0$)

$$\begin{aligned}\mathrm{ind}(P) &= \sum_{i\not\in I}\frac{d(u_i)^2|u_i|}{|u|}+\sum_{i\in I}\sum_{j=1}^{3}\frac{d(u_{i,j})^2|u_{i,j}|}{|u|}\\ &= \sum_{i\not\in I}\frac{d(u_i)^2|u_i|}{|u|}+\sum_{i\in I}\frac{\mathrm{ind}(u_{i,1}u_{i,2}u_{i,3})|u_i|}{|u|}\\ &\ge \sum_{i\not\in I}\frac{d(u_i)^2|u_i|}{|u|}+\sum_{i\in I}\frac{(d(u_i)^2+\varepsilon^3)|u_i|}{|u|}\\ &= \mathrm{ind}(S)+\varepsilon^3\frac{\sum_{i\in I}|u_i|}{|u|}\ge \mathrm{ind}(S)+\varepsilon^4\ .\end{aligned}$$

If P is not ε-regular, we repeat the splitting by part 2, and then iterate the splitting step until we get an ε-regular interval partition of u with t intervals. This always happens, without using any index increment bound. Since index does increase by at least ε^4 at each splitting and is bounded by 1, we terminate after at most $1/\varepsilon^4$ splittings and have the stated upper bound $t\le T_0$. □

Proof of the Axenovich–Person–Puzynina theorem

Let $\varepsilon\in(0,1)$ and $u\in\{0,1\}^n$ be given. We take the ε-regular interval partition $u=v_1v_2\dots v_t$ with $t\le 3^{1/\varepsilon^4}$ provided by Lemma 2 (used with $t_0=1$). For each ε-regular word v_i we take its large twin provided by Lemma 1 and concatenate them in the twin (u_1,u_2) in u. The rest of u goes of course in the bin u_3. How large is it? The total length on non-ε-regular v_is plus the total length of the bins of ε-regular v_is, which gives the bound

$$|u_3|<\varepsilon n+5\varepsilon n+3t\le 6\varepsilon n+3^{1+1/\varepsilon^4},\ n=1,2,\dots\ .$$

For large enough n, this is smaller than $7\varepsilon n$ and the proof is complete.

Concluding remarks

The article [2] investigates also more general scenarios with alphabets larger than binary and twins with more parts than two but here we restricted to the simplest but already intriguing case of binary words and two parts in a twin. It is shown in [2] (and it follow from the above displayed bound on $|u_3|$) that the εn in the theorem can be bounded by $\varepsilon n\ll n/(\log n/\log\log n)^{1/4}$, which is later in [2] improved to $\varepsilon n\ll n/(\log^{1/3}n/\log\log^{2/3}n)$ (I use $\ll$ as synonymous to $O(\cdot)$), but that it cannot be smaller than $\log n$.

We took the above proof from [2] but we did some simplifying (we specialize to the binary case and use simpler notion of ε-regularity of words) and rigorizing (we introduce back ceilings $\lceil\cdot\rceil$ and floors $\lfloor\cdot\rfloor$ omitted in [2], to be sure that the bounds are indeed correct; for example, the version of Lemma 1 in [2], Claim 11, asserts the bound $|u|-2|u_1|\le 5\varepsilon|u|$, which is suspicious for very small ε as for odd $|u|$ the left side cannot be smaller than 1).

We close with an interesting open problem, mentioned at the closing of [2] and then again in the survey [1, Open Problem 4.5]:

Does the A–P–P theorem hold for ternary words $u \in \{0, 1, 2\}^n$?

References

[1] M. Axenovich, Repetitions in graphs and sequences, preprint, July 2015.

[2] M. Axenovich, Y. Person and S. Puzynina, A regularity lemma and twins in words, J. Combin. Theory, Ser. A 120 (2013), 733–743.

[3] M. Klazar, A book on number theory, in preparation.