THE ROYAL SWEDISH ACADEMY OF SCIENCES

INSTITUT MITTAG-LEFFLER

Auravägen 17, SE-182 60 Djursholm, Sweden
Tel. +46 8 622 05 60 Fax. +46 8 622 05 89
info@mittag-leffler.se www.mittag-leffler.se

Avoiding tight twins in sequences via entropy compression

J. Grytczuk, J. Kozik and B. Zaleski

REPORT No. 50, 2013/2014, spring

ISSN 1103-467X
ISRN IML-R- -50-13/14- -SE+spring

AVOIDING TIGHT TWINS IN SEQUENCES VIA ENTROPY COMPRESSION

JAROSŁAW GRYTCZUK, JAKUB KOZIK, AND BARTOSZ ZALESKI

Abstract. A sequence S is called *twin-free* if no segment of S consists of two identical subsequences on disjoint index sets. This notion generalizes the famous *nonrepetitive* sequences of Thue. Using entropy compression technique we prove that there exists arbitrarily long twin-free sequences over 7-element set of symbols. Actually we prove a stronger statement that such sequences can be constructed from arbitrary lists of size 7. We also derive an extension of this result for *k-tuplet-free* sequences (no segment consists of k identical subsequences). The resulting bound on the number of symbols needed is of order $(1 + o(1))k$. The question whether a constant number of symbols is sufficient to avoid tuplets of any order in sequences remains open.

1. Introduction

Let $S = s_1s_2 \dots s_n$ be a finite sequence of symbols. Two subsequences $S_1 = s_{i_1}s_{i_2} \dots s_{i_k}$ and $S_2 = s_{j_1}s_{j_2} \dots s_{j_k}$ are called *twins* if they are identical and their sets of indices are disjoint, that is, if $s_{i_p} = s_{j_p}$ for all $p = 1, 2, \dots, k$, and $i_p \neq j_q$ for all $p, q = 1, 2, \dots, k$. This notion was introduced recently by Axenovich, Person, and Puzynina in [3]. They proved a beautiful result asserting that every binary sequence of length n contains a pair of twins of total length at least $n - o(n)$. The proof uses new regularity lemma for sequences.

In this paper we study the following problem inspired by their work. A pair of twins in a sequence S is called *tight* if the union of their index sets forms a full segment of integers. A sequence is called *twin-free* if it does not contain tight twins of any length. For instance, the sequence 23**121323**1 contains tight twins of the form 123, while 123132312 is twin-free (no block decomposes into a pair of identical subsequences). A special case of tight twins of the form $s_1s_2 \dots s_ns_1s_2 \dots s_n$ is called a *repetition*. The celebrated theorem of Thue [13] asserts that there exist arbitrarily long sequences without repetitions built of only three symbols. Does a similar result hold for twin-free sequences over finite number of symbols? We prove that the answer is positive in the following stronger sense.

J. Grytczuk acknowledges a support by the Institut Mittag-Leffler (Djursholm, Sweden) and the Polish National Science Center, decision nr DEC-2012/05/B/ST1/00652.

Theorem 1. *Let $n \geq 1$ be a natural number and let $L_1, L_2, \ldots, L_n$ be a sequence of arbitrary 7-element sets. Then there exists a twin-free sequence $S = s_1 s_2 \ldots s_n$ such that $s_i \in L_i$ for every $i = 1, 2, \ldots, n$.*

The proof uses a technique called *entropy compression* inspired by the algorithmic version of the Lovász Local Lemma due to Moser and Tardos [11]. When applied to sequences avoiding repetitions it gives a similar result with sets of size four (see [9]). We give a proof of Theorem 1 in Section 2. In Section 3 we extend the main result to sequences avoiding *tuplets* (segments consisting of a fixed number of identical subsequences). The paper is concluded with a final section containing open problems.

2. Proof of Theorem 1

We begin with presenting randomized algorithm that, with positive probability, generates a twin-free sequence of any given length.

2.1. Algorithm.

In the following algorithm we generate the sequence randomly from given lists, one element by one. If, by adding a new element to the sequence, we create a tight twin we find the index of the first element of the second twin and erase all elements starting with that index. For example if we obtain the sequence 123424121, where the last six digits form a tight twin, after the deletion we continue the algorithm with sequence 12342.

Algorithm 1. Creation of a twin-free sequence of length n.

$i \leftarrow 1$
while $i \leq n$ **do**
  $s_i \leftarrow$ *element of* $L_i \setminus \{s_{i-1}\}$ *chosen uniformly at random*
  if $s_1 s_2 \ldots s_i$ *is twin-free* **then**
    $i \leftarrow i + 1$
  else
    from all created twins choose the longest of length $2k$*, with* $k > 1$
    $b \leftarrow$ *index of the first element of the second twin*
    $i \leftarrow b$
  end
end

It is easy to observe that, as we never create a repetition of length 2, at least two elements remain after each of the erasures. Moreover, observe that the first two elements of a tight twin always belong to the first twin. Similarly the last two belong to the second twin.

We denote by A the length of the lists L_i, $i = 1, 2, \ldots, n$. We are going to prove that $A \geq 7$ is sufficient for the algorithm to succeed for every n with positive probability. Suppose that the algorithm does not stop before the M-th step (i.e. it makes at least M iteration of the main loop). In each iteration the algorithm makes some random decision. Each decision is the choice of one of $A - 1$ or A elements. Therefore the sequence of decisions can be described by a sequence $(s_1, \ldots, s_M)$ of numbers from $[A]$.

2.2. **Evaluation logs.** Fix a decision sequence $(s_1, \ldots, s_M)$ and assume that the algorithm did not succeed in M steps following this decision sequence. Clearly sequence $(s_1, \ldots, s_M)$ determines completely behavior of the algorithm in the first M steps. We are going to describe this run in another way.

Let $d_1 = 1$ and for $j = 2, 3, \ldots, M$ let d_j be the difference between variable i, during the evaluation of the algorithm, after step j and $(j-1)$. Sequence $(d_2, d_3, \ldots, d_M)$ consists of differences between the lengths of twin-free sequences generated during the run of the algorithm. We call such sequences $D = (d_1, \ldots, d_M)$ *supporting sequences*. The elements of sequence D are called *steps*. When $d_j = 1$ we say that the step is *positive* otherwise it is *negative*. Number $\sum_{i=1}^{M} d_i$ is called the *total difference* of d. Every supporting sequence satisfies the following two conditions

- $d_j \in \{\ldots, -3, -2, -1, 1\}$ for $j = 1, 2, \ldots, M$;
- $\sum_{j=1}^{k} d_j \geq 2$ for $k = 2, \ldots, M$.

The first one follows directly from the fact that we never create a repetition of length 2 and the second one from that we never remove the first two elements of a created tight twin.

Suppose that there are exactly t negative numbers in a supporting sequence $(d_1, d_2, \ldots, d_M)$. Let $-d$ be j-th such number, *pattern* of the j-th erasure is a sequence $P_j = (p_j^1, p_j^2, \ldots, p_j^{d+1})$ such that p_j^k is equal to 1 if the k-th removed element belonged to the first twin and -1 otherwise. For example, if at j-th step we create the following sequence $\ldots 12342415215$, where the last eight digits form a tight twin, then $d_j = -5$ and the pattern associated with it is $(-1, 1, 1, -1, -1, -1)$. Finally, with every such pattern P_j we associate the sequence E_j over $[A]$ called *base sequence* The length of that sequence is equal to the number of occurrences of 1 in the pattern (E_j may be the empty sequences), and each element describes some erased element of the first twin. An i-th element of the sequence E_j is the index of the corresponding element of the first twin within its corresponding list. In the above example, if all lists are equal to $\{0, 1, 2, 3, 4, 5\}$, we would observe $E_j = (2, 6)$. A quadruplet (D, P, E, C) is a *log* of length M if there exists a decision sequence $s_1, s_2, \ldots, s_M$ such that:

- the algorithm that follows this decision sequence does not stop before step M;
- D, P, E are corresponding sequence of differences, sequence of patterns, sequence of base sequences;
- C is the current twin-free sequence after M steps of the algorithm.

A triplet (D, P, E) that can be extended to a log of length M by appending some sequence C is called a *sketch* of length M. The following lemma is the first main ingredient of the proof.

Lemma 1. *Every log of length M corresponds to a unique decision sequence.*

Proof. Let

$$((d_1, d_2, \dots, d_M), (P_1, P_2, \dots, P_t), (E_1, E_2, \dots, E_t), C_M)$$

be a *log* with exactly t negative entries in D and $C_M = (c_1, c_2, \dots, c_l)$ for some integer $l > 0$. We prove the lemma by induction on M.

For $M = 1$ the assertion is obvious. Suppose the lemma is true for all numbers smaller than M.

If $d_M = 1$, then in the M-th step the element c_l is appended to C_M and thus $s_M = c_l$. Moreover

$$((d_1, \dots, d_{M-1}), (P_1, \dots, P_t), (E_1, \dots, E_t), (c_1, \dots, c_{l-1}))$$

is a *log* of length $M - 1$ and, by the induction hypothesis, we can retrieve the sequence $s_1, s_2, \dots, s_{M-1}$.

Suppose that $d_M < 0$. In that case $|d_M| + 1$ elements were erased from the sequence. Let there be exactly h elements in $E_t = (e_1, e_2, \dots, e_h)$ and r_1, r_2, r_h be the indices of "1" in P_t. Then, the length of a tight twin we created in step M is $2(|d_M| + 1 - h)$ and the sequence before the erasure was $(c_1, c_2, \dots, c_l, c_{l+1}, \dots, c_{l+|d_M|+1})$, where we first fill up the values of $c_l, c_{l+1}, \dots, c_{l+|d_M|+1}$ according to the pattern P_t. First, using consecutive elements of E_t, in those places where there is a "1" in the pattern, i.e. $c_{l+r_f} = e_f$ for $f = 1, 2, \dots, h$. Afterwards, sequentially using elements

$$(c_{l-(|d_M|+1-2h)+1}, c_{l-(|d_M|+1-2h)+2}, \dots, c_l, e_1, e_2, \dots, e_h),$$

in those places, where there is a 0 in P_t. Thus $s_M = c_{l+|d_M|+1} = e_h$ and from a *log* of length $M - 1$,

$$((d_1, \dots, d_{M-1}), (P_1, \dots, P_{t-1}), (E_1, \dots, E_{t-1}), (c_1, \dots, c_l, \dots, c_{l+|d_M|})),$$

we retrieve the sequence $s_1, s_2, \dots, s_{M-1}$, by the induction hypothesis. □

2.3. **Counting logs.** We are going to use the following observation which is a consequence of Proposition IV.5 from [7] (see also Corollary 11 in [6]).

Lemma 2. *Let ϕ be a function analytic at 0, having nonnegative Taylor coefficients, and such that $\phi(0) \neq 0$. Let $R \leq \infty$ be the radius of convergence of ϕ and assume that $\lim_{x \to R^-} \frac{x\phi'(x)}{\phi(x)} > 1$. Let $y(z)$ be the formal solution of the equation $y(z) = z\phi(y(z))$. Then for any $\tau_0 \in (0, R)$ for which $\frac{\tau_0 \phi'(\tau_0)}{\phi(\tau_0)} \neq 1$,*

$$[z^n]y(z) = o\left(\left(\frac{\phi(\tau_0)}{\tau_0}\right)^n\right).$$

The assertion of Theorem 1 follows from the following lemma.

Lemma 3. *If every list has 7 elements, then the number of logs of length M is $o(6^M)$.*

Proof. By the definition of the algorithm in every log (D, P, E, C) there are at most n elements in C. Observe also that if there are k elements in C, then the total difference of D is k. Obviously, the number of logs of length M is bounded from above by $A^n \sum_{k=2}^{n} L_M^k$, where L_M^k is the number of sketches (D, P, E) of length M with total difference k. For every supporting sequence $D = (d_0, d_1, \ldots, d_M)$ we have $d_0 = 1$, therefore we consider *trimmed supporting sequence* $(d_1, \ldots, d_M)$ instead. Every sequence $D = (d_1, \ldots, d_M)$ that can occur as a trimmed supporting sequence in some sketch satisfies the following conditions:

(1) $d_i \in \{1, -1, -2, -3, \ldots\}$, for every $i \in [M]$;
(2) $\sum_{i=1}^{k} d_i \geq 1$, for every $k \in [M]$;
(3) $\sum_{i=1}^{M} d_i \leq n$.

Every sequence of elements of $\{1, -1, -2, -3, \ldots\}$ which satisfies the above conditions will be called *admissible*. Sequences P and E that accompany D within a sketch can be seen as annotations of negative steps of D. Precisely, with every negative step of D with modulus k we associate a pattern of length $k+1$ and a sequence over $[A]$ of length l, where l is the number of occurrences of 1 in the pattern. Note that every pattern $(p_0, \ldots, p_k)$ which can occur in a sketch satisfies the following conditions:

(1) $p_0 = -1$;
(2) $\sum_{i=l}^{j} p_i > \sum_{i=l}^{k} p_i$, for $l < k$.

Once again since p_0 is always -1 we consider *trimmed pattern* $(p_1, \ldots, p_k)$ instead. Every pattern $(p_1, \ldots, p_k)$ that satisfies the second of the above conditions will be called *admissible*. We are going to count all admissible sequences in which every negative step is annotated by an admissible pattern and a base sequence of elements of $[A]$ of appropriate length. Clearly, not all of such structures correspond to some sketch. Let L_M^k be the number of such annotated sequences of length M with total difference k. Observe that $L_M^k \leq L_{M+1}^1$ (we can simply append $-k+1$ to D and annotate it with pattern of k "zeros" and the empty base sequence). Therefore, the total number of such annotated paths of length M is at most nL_{M+1}^1. Then, the total number of *logs* of length M is bounded from above by

$$A^n \cdot n \cdot L_{M+1}.$$

We say that sequence $(d_1, \ldots, d_M)$ *visits level* k at step l if $\sum_{i=1}^{l} d_i = k$. Every admissible sequence admits a canonical decomposition, known as *last passage decomposition*. If the sequence ends with 1 it has to be one element sequence. If it ends with $-d < 0$ then the sequence can be uniquely decomposed into $d+1$ subsequences $s_1, \ldots, s_{d+1}$ such that the j-th sequence consists of consecutive entries of D from the last visit to the level $j-1$ to the last visit to the level j. The original sequence d can be reconstructed from sequences $s_1, \ldots, s_{d+1}$ by concatenating them and appending $-d$. Note that subsequences $s_1, \ldots, s_d$ are admissible sequences of total difference 1. The same decomposition holds for annotated path with that modification that appending d to the concatenated sequences we have also to specify its annotation. Let F_k denote the number of admissible annotations for a negative step $-k$. The described decomposition implies that generating function $L(z) = \sum_{j=1}^{\infty} L_j^1 \cdot z^j$ satisfies the following functional equation

$$L(z) = z \left(1 + \sum_{k=1}^{\infty} L(z)^{k+1} \cdot F_k \right).$$

Then, if we denote by $F(x)$ generating function $\sum_{k=1}^{\infty} F_k x^k$, we get

$$L(z) = z(1 + L(z) \cdot F(L(z))). \tag{2.1}$$

Admissible annotation of a negative step $-k$ consists of a (trimmed) admissible pattern of length k and sequence of elements of $[A]$ of length l, where l is the number of occurrences of -1 in the pattern. It is convenient to analyze negated and reversed patterns instead of admissible ones. For pattern $(p_1, \ldots, p_k)$ inverted and negated pattern is $(-p_k, \ldots, -p_1)$. Admissibility condition for modified patterns becomes:

$$\sum_{i=l}^{j} p_i > 0, \text{ for every } l \in [k],$$

and the number of admissible annotations for negative step $-k$ is the number of such sequences of length k in which every -1 step is additionally annotated by a number from $[A]$. Observe that every modified pattern with $\sum_{j=1}^{k} = l$ can be decomposed into l Dyck paths. The i-th path of the decomposition is the sequence of consecutive elements of p from the fist to the last visit to the level i. The original sequence can be reconstructed from these paths by prepending each with 1 and then concatenating them. Let $C(y)$ be the generating function for Dyck paths such that $[y^k]C(y)$ is the number of Dyck paths of length $2y$. It is commonly known, and easy to derive, that $C(y) = \frac{1-\sqrt{1-4y}}{2y}$. Additionally every -1 step have to be annotated by some element of $[A]$. Altogether we get the following formula for function $F(x)$:

$$F(x) = \frac{1}{1 - x \cdot C(A \cdot x^2)} - 1.$$

Therefore equation (2.1) can be rewritten to

$$L(z) = z\phi(L(z)), \tag{2.2}$$

for function $\phi(x) = 1 + xF(x)$. For $A = 7$ the radius of convergence of $\phi(x)$ is $R = 1/(2\sqrt{7}) \approx 0.18898...$ and $\lim_{x \to R^-} \frac{x\phi'(x)}{\phi(x)} = +\infty$. We apply Lemma 2 with $\tau_0 = 0.18$. Then evaluating $x/\phi(x)$ at τ_0 we obtain some value $x_0 \approx 0.16845....$ In particular $x_0 > 1/6$ which implies that $L_M = o(6^M)$. Therefore, for fixed n, the number of logs of length M is $o(6^M)$ as well. $\square$

3. Avoiding tuplets in sequences

Let $k \geq 2$ be a fixed integer. By a *k-tuplet* we mean a sequence consisting of k identical subsequences whose index sets are pairwise disjoint. A sequence S is called *k-tuplet-free* if no segment of S is occupied by a k-tuplet.

Theorem 2. *For every k there exists number $f(k)$ such that for every sequence of lists, each of length $f(k)$, it is possible to choose k-tuplet-free sequence from these lists.*

Proof. We describe briefly how the proof of Theorem 1 can be adapted to this problem. To keep things simple we modify the Algorithm 1 so that it removes the whole k-tuplet when it is produced and does not explicitly avoid occurrences of short k-tuplets (i.e. the next element is chosen from the full list, without removing previous symbol). Just like before a log is a quadruple (D, P, E, C). Supporting sequence D is a sequence of numbers from $\{1, -k+1, -2k+1, -3k+1, \ldots\}$. P is a sequence of patterns i.e. sequences over $[k]$, such that the length of the i-th pattern is $|d_j|+1$, where d_j is the value of the i-th negative step of D. Finally each pattern from P of the length $l \cdot k$ has corresponding base sequence over $[A]$ of length l in E. Therefore each negative step d_j of value $-l \cdot k + 1$ can be annotated in less than $F_{kl} = A^l k^{lk}$ ways. The generating function of this sequence $F(x) = \sum_{l=1}^{\infty} A^l (xk)^{kl}$ is $F(x) = \frac{1}{1-A(xk)^k} - 1$. Just like in the proof of Theorem 1 the number of possible logs of length M is $O([z^M]L(z))$ for function $L(z)$ defined by the following functional equation

$$L(z) = z(1 + \sum_{l=1}^{\infty} L(z)^{kl} F_{kl}).$$

The equation can be rewritten as

$$L(z) = z\phi(L(z)),$$

where $\phi(x) = 1 + F(x)$. The radius of convergence of $\phi(x)$ is $R = A^{-1/k}k^{-1}$ and $\lim_{x \to R^-} \frac{x\phi'(x)}{\phi(x)} = \infty$. Therefore ϕ satisfies assumptions of Lemma 2. For $A = 2k$ we can choose $\tau_0 = 2/(3k)$. Then we obtain

$$\phi(\tau_0)/\tau_0 = \frac{3}{2\left(1 - 2k\left(\frac{2}{3}\right)^k\right)} k,$$

which for large enough k is smaller than $2k$. Therefore the number of logs of large length is smaller than the number of input sequences. That implies that some input sequences generates k-tuplet-free sequences of desired length. □

More careful estimations can reduce alphabet from $2k$ to $k(1 + o(1))$ but the method does not seem to be able to provide sublinear bound.

4. Final remarks

The first natural question concerns optimality of the bound given in Theorem 1. We suspect that it is not best possible and pose the following conjecture.

Conjecture 1. *Let $n \geq 1$ be a natural number and let $L_1, L_2, \ldots, L_n$ be a sequence of 5-element sets. There exists a twin-free sequence $S = s_1, s_2, \ldots, s_n$, such that $s_i \in L_i$ for every $i \in [n]$.*

One way to achieve a better upper bound might be using a bit different algorithm to generate our sequence. Instead of avoiding repetitions of length 2 one could avoid also repetitions of length 4. Unfortunately in this case both the lower and upper bound on the number of possible sequences generated by the algorithm become much more complicated. Interestingly, by means of simple case analysis, we were able to show the lower bound of 4. But, due to performed simulations, we believe that it is not the proper value.

Another intriguing question concerns optimality of Theorem 2. Actually, one expects that it should be easier to avoid k-tuplets for larger k and that perhaps $f(k)$ becomes constant for sufficiently large k. The following conjecture offers the strongest statement in this direction.

Conjecture 2. *There exist arbitrarily long binary sequences avoiding k-tuplets for all sufficiently large k.*

References

[1] J.-P. Allouche, J. Shallit. Automatic Sequences. Theory, Applications, Generalizations, Cambridge University Press, Cambridge, 2003.
[2] N. Alon, J. Grytczuk, M. Hałuszczak, O. Riordan. Non-repetitive colorings of graphs. Random Structures Algorithms 21 (2002) 336–346.
[3] M. Axenovich, Y. Person, S. Puzynina, Regularity lemma and twins in sequences, J. Combin. Theory, Ser. A 120 (2013) 733–743.
[4] J. Berstel, Axel Thue's papers on repetitions in words: a translation, Publications du LaCIM, vol. 20, Université du Québec a Montréal, 1995.
[5] J. Berstel, D. Perrin, The origins of combinatorics on words, Europ. J. Combin. 28 (2007) 996–1022.
[6] V. Dujmović, G. Joret, J. Kozik, D. R. Wood. Nonrepetitive colouring via entropy compression, Combinatorica (to appear).
[7] P. Flajolet, R. Sedgewick, Analytic combinatorics, Cambridge University Press, 2009.
[8] J. Grytczuk, Thue type problems for graphs, points, and numbers. Discrete Math. 308 (2008) 4419–4429.
[9] J. Grytczuk, J. Kozik, P. Micek, New approach to nonrepetitive sequences. Random Structures Algorithms, DOI 10.1002/rsa.20411.
[10] M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading, MA, 1983.
[11] R. Moser, G. Tardos, A constructive proof of the general Lovász local lemma, J. ACM, 57 (2) (2010).
[12] J. Nešetřil, P. Ossona de Mendez, Sparsity, Springer, 2012.
[13] A. Thue, Über unendliche Zeichenreichen, Norske Vid. Selsk. Skr., I Mat. Nat. Kl., Christiania 7 (1906) 1–22.

Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-611 Warszawa, Poland
E-mail address: j.grytczuk@mini.pw.edu.pl

Faculty of Mathematics and Computer Science, Jagiellonian University, 30-348 Kraków, Poland
E-mail address: kozik@tcs.uj.edu.pl

Faculty of Mathematics and Computer Science, Adam Mickiewicz University, 61-614 Poznań, Poland