ssdnm
środowiskowe
studia doktoranckie
z nauk matematycznych

Bartosz Zaleski

Uniwersytet A. Mickiewicza w Poznaniu

Avoiding tight twins in sequences by entropy compression

Praca semestralna nr 3

(semestr letni 2012/13)

Opiekun pracy: Jarosław Grytczuk

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

AVOIDING TIGHT TWINS IN SEQUENCES BY ENTROPY COMPRESSION

JAROSŁAW GRYTCZUK, JAKUB KOZIK, BARTOSZ ZALESKI

Abstract. A sequence $T = t_1, t_2, \ldots, t_{2k}$ is called a tight twin if it can be divided into two identical disjoint subsequences. We say that a sequence S is twin-free if it has no tight twins as subsequences. This notion is a natural generalisation of a well known problem of nonrepetitive sequences posed by Thue in [5]. It is a well known fact that there are arbitrarily long nonrepetitive sequences over a 3-element set of symbols. Recently Grytczuk, Przybyło and Zhu [6] showed, using Lefthanded Local Lemma, that a similar result holds if we choose elements from 4-element lists attached to each position in the sequence. In [1] Grytczuk, Kozik and Micek proposed a different approach to this problem utilizing so called entropy compression argument. This powerful argument allows, by means of simple counting, to obtain results comparable to such elaborate tool as Lefthanded Local Lemma. In this paper we use entropy compression to show that a 7-element lists of symbols are enough to produce arbitrarily long twin-free sequences.

1. Introduction

A subsequence $T = x_1, x_2, \ldots, x_{2k}$ of consecutive elements of a sequence S is called a tight twin of length $2k$ if there exists a partition of elements of T into disjoint subsequences $T_1 = x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ and $T_2 = x_{j_1}, x_{j_2}, \ldots, x_{j_k}$, such that $x_{il} = x_{j_l}$ for every $l \in [k]$. We call the first subsequence by *the first twin* and the second one by *the second twin.* For instance, following sequences are tight twins

- 1234**1234**
- 12**1**3**23**4**4**
- 134**1**2**342**.

A sequence S is twin-free if it does not contain a tight twin of any length $2k > 0$.

The notion of twin-free sequences arises naturally as a generalisation of nonrepetitive sequences in which we only avoid twins formed by consecutive blocks of elements. It is easy to see that it is not possible to construct a nonrepetitive sequence of length higher than 3 over 2-element set of symbols. It was shown by Thue [5] that 3-elementary alphabet is enough.

Inspired by the algorithmic proof of Lovász Local Lemma [4] Grytczuk et al. [1] showed slightly weaker upper bound for list version of the problem. In the list setting every element of a sequence must be chosen from a different list of elements. The authors were able to show that lists of size at least 4 are enough. Remarkably, the same bound was earlier proved in [6] by the application of more elaborate technique - Lefthanded Local Lemma. It was also conjectured that lists of size 3 suffice.

In this paper we focus on showing yet another application of the method used in papers mentioned above. We prove the following theorem.

Theorem 1.1. *Let $n \geq 1$ be a natural number and let $L_1, L_2, \ldots, L_n$ be a sequence of 7-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 $i \in [n]$.*

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

2. 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 a following sequence 123424121, where the last six digits form a tight twin, after the deletion we continue the algorithm with sequence 12342.

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.

Algorithm 1: Creation of twin-free sequence of length n

```
$i \leftarrow 1$
while $i \leq n$ do
    $s_i \leftarrow$ element of $L_i \backslash \{s_{i-1}\}$ chosen uniformly at random
    if $s_1, s_2, \ldots, s_i$ is twin free then
        $i \leftarrow i + 1$
    else
        there exists exactly one tight twin of length $2k$ for some $k > 1$
        $b \leftarrow$ index of the first element of the second twin
        $i \leftarrow b$
    end
end
```

Proof of Theorem 1.1. Let $n \geq 1$ be a natural number and suppose by contradiction that the algorithm can not create the twin-free sequence for such n. Moreover suppose that each list L_i is of size A

Let M be a sufficiently large integer and run the algorithm for M steps. We double count the number of possible such executions, i.e. the number of all possible sequences of random values produced by the algorithm during the first M steps. Trivially, as we remove at most one element from each list, there are at least $(A-1)^M$ such sequences.

Fix an evaluation of s_1, s_2, s_M. Let $d_1 = 1$ and for $j = 2, 3, \dots, M$ d_j be the difference between variable i after step j-th and $(j-1)$-th. So $d_2, d_3, \dots, d_M$ is the sequence of differences between the lengths of twin-free sequences generated during the running of the algorithm. This sequence satisfies following two conditions

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

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. Let

Suppose that there are exactly t negative numbers in $(d_1, d_2, \dots, d_M)$. Let d be j-th such number, by $P_j = (p_j^1, p_j^2, \dots, p_j^{d+1})$ we denote the binary sequence of length $|d| + 1$ describing the pattern of the erased part of a tight twin, i.e. p_j^k is equal to 1 if the k-th removed element belonged to the first twin and 0 otherwise. For example, if at j-th step we create a following sequence $\dots 12342415215$, where the last six digits form a tight twin, then $d_j = -5$ and the pattern associated with it is $(0, 1, 1, 0, 0, 0)$. Finally, for every such pattern P_j we associate a sequence E_j of elements that belongs to the first twin in the erased part of the sequence. If there were no such elements (i.e. we created a repetition and removed only elements of the second twin) we put $E_j = ()$. In the example above $E_j = (1, 5)$.

A quadruplet (D, P, E, C) is a *log* of length M if there is an evaluation of $s_1, s_2, \dots, s_M$ such that D is the corresponding sequence of differences, P sequence of patterns, E sequence of sequences of erased elements from each first twin and C is the twin-free sequence created by the algorithm after M steps. A triplet (D, P, E) satisfying those constraints is called a *sketch* of length M. We show that having a *log*, one can retrieve the whole sequence $(s_1, s_2, \dots, s_M)$.

Lemma 2.1. *Every log corresponds to a unique evaluation of* $(s_1, s_2, \dots, s_M)$.

Proof of Lemma. 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 $M = 1$ it 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

$$((d_1, d_2, \dots, d_{M-1}), (P_1, P_2, \dots, P_t), (E_1, E_2, \dots, E_t), (c_1, c_2, \dots, c_{l-1}))$$

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 $f = 1, 2, \ldots, h$. Afterwards, sequentially using elements $(c_{l-(|d_M|+1-2h)+1}, c_{l-(|d_M|+1-2h)+2}, \ldots, c_l, e_1, e_2, \ldots, e_h)$, 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 $M-1$

$$((d_1, d_2, \ldots, d_{M-1}), (P_1, P_2, \ldots, P_{t-1}), (E_1, E_2, \ldots, E_{t-1}), (c_1, c_2, \ldots, c_l, \ldots, c_{l+|d_M|})),$$

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

As we assumed that the algorithm can not produce a twin-free sequence then in every *log* (D, P, E, C) of length M there are at most $n-1$ elements in C. Observe that if there are k elements in C, then

$$\sum_{j=1}^{M} d_j \quad = \quad k. \tag{1}$$

Obviously the number of possible *logs* is bounded from above by the $A^n \sum_{k=2}^{n-1} L_M^k$, where L_M^k is the number of sketches (D, P, E) of length M satisfying (1). Moreover, observe that $L_M^k \le L_{M+1}^2$ (we simply append $-k+1$ to D, sequence of $(k-2)$ zeros to P and an empty sequence to E) and $L_{M+1}^2 = L_M$, where L_M is the number of *sketches* with the first "1" removed from D, satisfying 1 with $k = 2$ and with all possible sequences of elements of appropriate length allowed in E. We call such structure a *restricted sketch.* Thus the total number of *logs* of length M and consequently sequences $s_1, s_2, \ldots, s_M$ is bounded from above by

$$A^n \cdot n \cdot L_M.$$

To estimate the number L_M we need some facts about generating functions. We call a generating function $f(z)$ with positive radius of convergence algebraic if there exists a nonconstant polynomial $W(z, f) \in \mathbb{C}[z, f]$ (defining polynomial) such that $W(z, f)$ is constantly zero within the disc of convergence of $f(z)$. Trivially, the coefficients f_n for $n = 0, 1, 2, \ldots$ of a generating function $f(z) = \sum_{n=0}^{\infty} f_n z^n$ with a radius of convergence strictly greater then λ satisfy $f_n = o(\lambda^{-n})$. We need following well known fact about algebraic generating functions.

Fact ([3]). *Let $f(z) = \sum_{n=0}^{\infty} f_n z^n$ be a nonpolynomial algebraic generating function with defining polynomial $W(z, t)$. Then the radius of convergence of $t(z)$ is one of the roots of the discriminant of $W(z, t)$ with respect to the variable t.*

Let $C_{0-1}(z)$ be the ordinary generating function in which $[z^k]C_{0-1}(z)$ counts the number of sequences $(x_1, x_2, \ldots, x_k)$, where

- $x_i \in \{0, 1\}$ for $i = 1, 2, \ldots, k$,
- $\sum_{i=1}^{k} x_i = 1$,
- $\sum_{i=1}^{j} x_i \ge 1$ for $j = 1, 2, \ldots k$.

We call such sequence a *binary zig-zag.* Every such sequence is either a singleton (1) or can be uniquely decomposed into l subsequent *binary zig-zags*, where

- the j-th component is a subsequence starting with the last such h that $\sum_{i=1}^{h} x_i = j-1$ and ending with the last such g that $\sum_{i=1}^{g} x_i = j$
- l is the number of " -1" at the end of the sequence, increased by 1.

Thus, $C_{0-1}(z)$ satisfy the following functional equation:

$$C_{0-1}(z) = z + z(1 + zC_{0-1}(z) + z^2 C_{0-1}^2(z) + \dots), \tag{2}$$

and consequently $C_{0-1}(z) = \frac{z}{1-zC_{0-1}(z)}$. Next, let $P(z)$ be the ordinary generating function in which $[z^k]P(z)$ counts the number of possible patterns of length k that may appear in *restricted sketches* with each of "1" additionally annotated with an element of an appropriate list L_i. Observe that each of those patterns P_j may be concatenated with a string of some number of leading "1" to obtain a pattern that describes a full tight twin. Now, we look at P_j from the end. We treat every "0" as "1" and every "1" as "-1". In this way we obtain a sequence that satisfy (2) and (2) and (2) with some number l instead of 1. We call this sequence a *binary extended zig-zag*. Observe that each such sequence can be decomposed into $l+1$ *binary zig-zags* in a similar manner as before. The number of "1's" that need to be annotated in the restricted sketch of length k with corresponding *binary extended zig-zag* that in (2) sums to l is exactly $\frac{k-l}{2}$. Thus the generating function $P(z)$ can be written as

$$P(z) = z\Big(\frac{C_{0-1}(z\sqrt{A})}{\sqrt{A}} + \frac{C_{0-1}^2(z\sqrt{A})}{\sqrt{A}^2} + \dots\Big), \tag{3}$$

or $P(z) = z\frac{1}{1-\frac{C_{0-1}(z\sqrt{A})}{\sqrt{A}}} - z$. Finally, let $L(z) = \sum_{n=0}^{\infty} L_n z^n$ be the generating function that counts the number of *restricted sketches.* We claim that function $L(z)$ satisfies

$$L(z) = x + xP(L(z)). \tag{4}$$

Indeed, again for every *restricted sketch* either D consists of a single "1", or it can be annotated with a *binary extended zig-zag* of length $|d_M|+1$ and decomposed into $|d_M|+1$ *restricted sketches* of total length $M-1$. But the number of such *extended zig-zags* is exactly $[x^{|d_M|+1}]P(x)$. Thus, for any integer $k > 1$,

$$\sum_{i=1}^{k-1}[L^i]P(L) \sum_{\substack{a_1+\dots+a_i=k-1 \\ a_1,\dots,a_i>0}} [x^{a_1}]L(x)\cdot[x^{a_2}]L(x)\cdot\ldots\cdot[x^{a_i}]L(x)$$

is equal to $[x^{k-1}P(L(x))]$ and counts the number of *restricted sketches* of length k. Putting together relations (2), (3) and (4) we obtain the defining polynomial for $L(x)$ to be

$$W(z,L) = x^2 + (2x + 2x^2 + Ax^2)z + (-1 - 3x - 2Ax - 2x^2)z^2 + (1 + A + 2x + x^2)z^3$$

and the discriminant of this polynomial is

$$x^4 - 2Ax^5 - 6x^6 + A^2x^6 - 8x^7 + 6Ax^7 + 8A^2x^7 - 3x^8 + 4Ax^8 - 20A^2x^8 - 4A^3x^8.$$

For $A = 7$ it has only one positive root, namely $x_0 \approx 0.171444 > 6^{-1}$. Thus we can take any $x_0^{-1} < \alpha < 6$ and get that $L_M = o(\alpha^M)$. Combining two bounds on the number of possible *logs* we have that

$$6^M \leq 6^n \cdot n \cdot o(\alpha^M),$$

which for sufficiently large M and fixed n yields a contradiction. □

3. Remarks

The result we presented in the previous section gives an upper bound on the sizes of lists from which we build a twin-free sequence. We suspect that it is not the optimal value and pose the following conjecture:

Conjecture. *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 the lower 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.

References

[1] J. Grytczuk, J. Kozik, P. Micek, *New approach to nonrepetitive sequences*, Random Struct Algorithms **42** (2012), 214–225.

[2] M. Molloy, *Cores in random hypergraphs and boolean formulas*, Random Structures and Algorithms **27** (2005), 124–135.

[3] P. Flajolet, R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009.

[4] R. A. Moser, G. Tardos, *A constructive proof of the general lovász local lemma*, J ACM **57** (2010).

[5] A. Thue, *Über unendliche zeichenreichen*, Norske Vid Selsk Skr I Mat Nat Kl Christiania (1906), 122.

[6] J. Grytczuk, J. Przybyło, N. Wormald, *Nonrepetitive list colorings of paths*, Random Struct Algorithms **38** (2011), 162173.

Faculty of Mathematics and Computer Science, Adam Mickiewicz University, ul. Umultowska 87, 61-614 Poznań, Poland